Stochastic Approximation Beyond Gradient for Signal Processing and Machine Learning

Aymeric Dieuleveut, Gersende Fort, Éric Moulines, Hoi-To Wai

Stochastic Approximation Beyond Gradient for Signal Processing and Machine Learning

Aymeric Dieuleveut, Gersende Fort, Eric Moulines, Hoi-To Wai, *Member, IEEE*

***Abstract*—Stochastic Approximation (SA) is a classical algorithm that has had since the early days a huge impact on signal processing, and nowadays on machine learning, due to the necessity to deal with a large amount of data observed with uncertainties. An exemplar special case of SA pertains to the popular stochastic (sub)gradient algorithm which is the working horse behind many important applications. A lesser-known fact is that the SA scheme also extends to non-stochastic-gradient algorithms such as compressed stochastic gradient, stochastic expectation-maximization, and a number of reinforcement learning algorithms. The aim of this article is to overview and introduce the non-stochastic-gradient perspectives of SA to the signal processing and machine learning audiences through presenting a design guideline of SA algorithms backed by theories. Our central theme is to propose a general framework that unifies existing theories of SA, including its non-asymptotic and asymptotic convergence results, and demonstrate their applications on popular non-stochastic-gradient algorithms. We build our analysis framework based on classes of Lyapunov functions that satisfy a variety of mild conditions. We draw connections between non-stochastic-gradient algorithms and scenarios when the Lyapunov function is smooth, convex, or strongly convex. Using the said framework, we illustrate the convergence properties of the non-stochastic-gradient algorithms using concrete examples. Extensions to the emerging variance reduction techniques for improved sample complexity will also be discussed.**

***Index Terms*—stochastic approximation, convergence analysis, compressed stochastic gradient, expectation maximization**

I. Introduction

Stochastic Approximation (SA) is a classical iterative algorithm that has a long history of over 70 years [1], [2]. The goal of the SA scheme is to determine the roots of a nonlinear system when the mean field cannot be explicitly computed but a random oracle exists. The SA scheme has had since the early days a huge impact in signal processing and automatic control: the first applications focused on the adaptive identification of systems [3]–[6]. Recently, the spectrum of use of SA schemes has widened considerably with the applications to statistical machine learning; see [7]–[9].

An extensive literature on stochastic optimization is devoted to the stochastic (sub)gradient (SG) algorithm, which is by far the most popular application of the SA scheme, see *e.g.,* [10]–[12] and the references therein. The stochastic gradient algorithms are characterized by having an update recursion featuring an *unbiased* mean field which is the gradient of a loss function to be minimized. However, a lesser known fact is that SA scheme also includes *non-stochastic-gradient (non-SG)* algorithms whose stochastic oracles are not the gradients of any function and whose oracles are possibly *biased* even in the asymptotic sense. Recently, these *non-SG* algorithms have gained attention in many scenarios of Signal Processing (SP) and Machine Learning (ML). Classical examples include preconditioned least mean square for linear system identification [13]; other more subtle examples include natural gradient methods [14], [15], online blind source separation [16]–[19], straight-through compressed gradient estimator [20], [21] and randomized coordinate descent algorithm [22].

The SA schemes that are non-stochastic-gradient algorithms appear quite naturally in modern statistical learning. An example is the determination of stationary points appearing in the stochastic versions of the Majorize-Minimization (MM) methods [23], [24] such as the SA versions of the Expectation-Maximization (EM) algorithm [25]–[27]. Another example is the algorithms used in reinforcement learning, such as TD-learning and Q-learning, in which the iterative mapping is deduced from the Bellman operator [27]–[30].

To illustrate the application of the SA scheme as non-stochastic-gradient algorithms, we may take a closer look at the family of stochastic EM algorithms introduced in [25]. While EM can be derived by the majorization-minimization method, a powerful perspective presented in [25] is to view the expectation step ('E-step') as fixed-point iterations in *sufficient statistics* space (the s-space). This perspective allows us to build highly efficient stochastic EM algorithms for streaming data and big data [26], [27], [31]–[33]. The challenge in their analysis is that this fixed point map in s-space is not the stochastic gradient map of *any* function. Therefore, the stochastic EM algorithms are in fact *not* SG algorithms. The convergence analysis of these algorithms to cope with the maximum likelihood estimation will be analyzed using the SA scheme, which includes the non-SG algorithms.

There are many excellent overview articles or books on SA scheme. We note that classical books such as [5], [13], [34] focused on asymptotic convergence for *unbiased* SA under a set of restrictive stepsize conditions. Along the line of applications on adaptive filtering, [35] presents finite-time analysis on the family of adaptive filtering algorithms such as least mean squares, recursive least squares, etc. The recent articles [10]–[12] are devoted to the *gradient-based* SA schemes. While they provide a modern treatment on the convergence of SA schemes, the discussions are limited to stochastic gradient algorithms. The recent book [36] includes results that are

AD and EM are with Ecole Polytechnique, CMAP, UMR 7641, France. GF is with Institut de Mathématiques de Toulouse, UMR5216, Université de Toulouse, CNRS; UPS, F-31062 Toulouse Cedex 9, France. HTW is with CUHK, Hong Kong. Work partly supported by the *Fondation Simone et Cino Del Duca, Institut de France*, ANR under the program MaSDOL-19-CE23-0017-01, ANR-19-CHIA-SCAI-002, HKRGC Project 24203520, Hi!Paris *FLAG* project, and been carried out under the auspices of Lagrange research Center for Mathematics and Calculus.

applicable to non-SG algorithms, but is otherwise focused on SA schemes matched to find the roots of the gradient of a strongly convex objective function.

Most of the existing overviews on the convergence for SA schemes are not comprehensive when it comes to discussing the results for non-stochastic-gradient algorithms, or they are limited to asymptotic convergence for algorithms with decreasing step sizes sometimes accompanied by limit distributions (in favorable cases). A possible reason behind this is the lack of a *proper* Lyapunov function to set the convergence analysis framework, and the potential bias may destabilize the SA recursion. To this end, there are no convergence results (or only in certain cases) for the beyond-gradient-SA scheme in the literature at the level of generality of SG algorithms such as [12]. Also missing is a principled guide to designing or improving algorithms for SP and ML that do not originate from Stochastic Gradient.

This article fills a gap in the literature by proposing a general framework that summarizes recent advances in the theories of the SA scheme, with emphasis on the recent applications to SP and ML. Our aim is to provide an up-to-date overview of the classical SA scheme for researchers working in the fields of SP and/or ML. We shall cover results from the basic insights behind the SA scheme, to standard analysis on convergence in expectation, to the advanced analysis with almost-sure convergence. Our results will handle the challenging settings of generic non-SG (possibly biased) algorithms.

In the **first part** of this paper, the SA scheme is introduced as a root-finding algorithm using a stochastic oracle, designed through Euler discretization of an ordinary differential equation flow (see Section II-A). Then, in Section II-B, examples of SA schemes will be presented with a focus on non-SG algorithms such as compressed SG algorithms, stochastic expectation maximization, and policy evaluation via temporal difference learning. Section III-A discusses the general assumptions required for the convergence of the SA scheme, where we formalize a set of conditions for a *proper Lyapunov* function design with respect to the SA scheme that can be potentially biased. Section III-B shows how these conditions are satisfied in the applications listed. Readers are recommended to read this part first to get themselves familiar with the basics of SA scheme.

We next present the general convergence theories for SA in two different favors. The **second part** is devoted to non-asymptotic convergence bounds of the SA scheme that focuses on the *expected convergence* towards the root(s) of a nonlinear system in a finite number of iterations. Section IV-A describes a unified result on finite-time bounds and sample complexity. One of the features of our study is to investigate the effects of the presence of bias in the stochastic oracle, a situation that often arises in applications, e.g., when the stochastic oracle uses compression or quantization, or uses Monte Carlo methods - importance sampling or Markov chain Monte Carlo that are inherently biased - see [27], [37], [38]. In Section IV-B, we illustrate the application of these results using the examples introduced in Section II-B and we discuss the obtained results with the state of the art.

The **third part** is devoted to the almost-sure convergence of the sequence of iterates generated by the SA scheme towards the root(s) of a nonlinear system when the number of iterations goes to infinity. Section V-A gives a brief overview of the theory of Stochastic Approximation with decreasing step size. In Section V-B we consider the case where the approximation bias vanishes asymptotically. We use the *Ordinary Differential Equation* (ODE) method, which relates the almost sure limit sets of the stochastic approximation process to the limit sets of the flow of an ODE. We also establish almost-sure convergence results under the same assumptions as for the non-asymptotic approaches. In Section V-C, we extend these results to the case where the bias in the approximation does not vanish asymptotically. This is a case that has been much less studied in the literature. While we use [39], [40], which deals with stochastic-gradient-type SA schemes, the results presented in this section are original.

Finally, in the **fourth part** we review a recent advance in SA: we discuss in particular the *Stochastic Path-Integrated Differential EstimatoR* method (SPIDER) originally introduced for stochastic gradient algorithms [41], [42] and then extended to EM [33], [43]. We provide here an algorithmic description and a non-asymptotic convergence analysis for a general SA scheme (see also [38]).

We refer the readers to Fig. 1 for a guide to navigate through this overview article.

Notations. $\mathbb{N}$ is the set of the non-negative integers, $\mathbb{R}$ is the set of the real numbers, $\mathbb{R}_+$ is the set of the non-negative real numbers, and $\bar{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{+\infty\}$ is its completed version. When necessary, we use the convention $0^{-1} = \infty$ and $\infty \times 0 = 0$. The set of the minimizers and maximizers of a function F is denoted by $\arg\min$ (resp. $\arg\max$) when it is a singleton, and Argmin (resp. Argmax) otherwise. For two real numbers x and y, $x \wedge y$ (resp. $x \vee y$) denotes the minimum (resp. maximum) of x and y. $\lfloor x \rfloor$ is the floor of x.

For a vector or matrix $\mathbf{D}$, $\mathbf{D}^\top$ is the transpose of $\mathbf{D}$. Vectors are column-vectors. For two vectors $\boldsymbol{w}, \boldsymbol{w}'$ in $\mathbb{R}^d$, $\langle \boldsymbol{w} \,|\, \boldsymbol{w}' \rangle := \boldsymbol{w}^\top \boldsymbol{w}'$ is the dot product of $\boldsymbol{w}$ and $\boldsymbol{w}'$. We set $\|\boldsymbol{w}\| := \sqrt{\langle \boldsymbol{w} \,|\, \boldsymbol{w} \rangle}$. For a positive-definite matrix $\mathbf{D}$, we denote by $\|\boldsymbol{w}\|_{\mathbf{D}} := \sqrt{\boldsymbol{w}^\top \mathbf{D} \boldsymbol{w}}$ the norm associated to the scalar product induced by $\mathbf{D}$.

For a differentiable function $\boldsymbol{h}$, its gradient is denoted by $\nabla \boldsymbol{h}$. When $\boldsymbol{H}$ is a function of two variables $(\boldsymbol{w}, \mathbf{x}) \mapsto \boldsymbol{H}(\boldsymbol{w}, \mathbf{x})$, we will write $\mathsf{D}_{ij} \boldsymbol{H}(\boldsymbol{w}, \mathbf{x})$ for the i-th derivative *w.r.t.* the variable $\boldsymbol{w}$ and the j-th derivative *w.r.t.* the variable $\mathbf{x}$ of the function $\boldsymbol{H}$, evaluated at $(\boldsymbol{w}, \mathbf{x})$.

All the random variables are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. $\mathbb{E}$ is the expectation associated to the probability $\mathbb{P}$. When $\mathcal{F}$ is a filtration, $\mathbb{E}^{\mathcal{F}}\big[\boldsymbol{X}\big]$ is the expectation of the random variable $\boldsymbol{X}$ conditionally to $\mathcal{F}$. When $\mathcal{F}$ is equal to $\sigma(\mathbf{U})$, the σ-field generated by the random variable $\mathbf{U}$, we will write $\mathbb{E}^{\mathbf{U}}\big[\boldsymbol{X}\big]$. The limit set of a sequence is denoted by $\mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$: $\mathbf{w}_\star \in \mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$ if $\lim_{k\to\infty} \mathbf{w}_{n_k} = \mathbf{w}_\star$ for some subsequence $\{n_k,\ k \in \mathbb{N}\}$ such that $\lim_{k\to\infty} n_k = +\infty$.

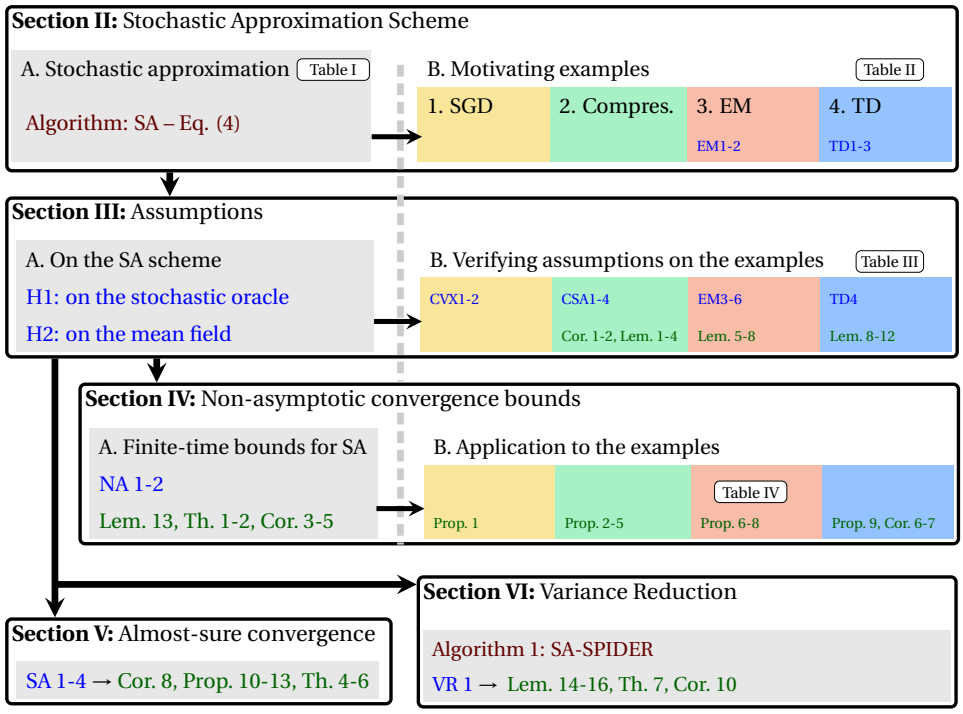


Fig. 1: Guide for navigating through the content in this overview article. **Outline: Section II** introduces the central SA algorithm and its derivation on four motivating examples, and Section III the two main assumptions on the oracle and the field for the analysis to hold, with corresponding derivation on examples. Theoretical results are given in **Section IV**, **Section V**, and **Section VI**. **Dependencies and reading guide:** Gray background indicates generic results on the SA scheme, and colored backgrounds indicate their application to the four examples. In Sections II to IV, subsection A deals with the generic scheme and subsection B.1 to B.4 with the four examples. Black font indicates the sections, blue font indicates assumptions that are made for the analysis; green font corresponds to the results (lemmas, propositions, corollaries and theorems), and red font to algorithms. Arrows indicate dependencies: readers may choose to read only the generic results on the SA scheme, together with some of the examples.

II. Stochastic Approximation Scheme

A. Stochastic Approximation

Stochastic Approximation (SA) is a class of stochastic algorithms aiming at a solution to the root-finding problem:

$$\text{find} \quad \boldsymbol{w}^\star \in \mathbb{R}^d \quad \text{such that} \quad \boldsymbol{h}(\boldsymbol{w}^\star) = \boldsymbol{0} \,, \tag{1}$$

where $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^d$, $d \in \mathbb{N}$, is known as the mean field function. Problem (1) is motivated by many tasks in SP and ML, such as convex or non-convex optimization [12], statistical estimation [31], policy evaluation [29], etc., some of them will be described later in Section II-B.

Solving (1) is challenging when *only* stochastic estimates of the map $\boldsymbol{h}(\cdot)$ are available. We follow the ordinary differential equation (ODE) approach introduced in [3], [4]. This approach consists in identifying the limit points of the SA sequence with a sub-class of invariant sets of the ODE flow (see Section V for precise definitions)

$$\mathrm{d}\boldsymbol{w}/\mathrm{d}t = \boldsymbol{h}(\boldsymbol{w}(t)) \,. \tag{2}$$

In particular, the equilibrium points of the ODE (*i.e.,* the points $\boldsymbol{w}_\star$ satisfying $\boldsymbol{h}(\boldsymbol{w}_\star) = 0$) are the solutions of (1).

To obtain a discrete-time algorithm, the common trick is to discretize the continuous time algorithm (2) through the Euler's scheme. Set $\boldsymbol{w}_k \equiv \boldsymbol{w}(k\gamma)$, with a sufficiently small discretization parameter $\gamma > 0$, the time-derivative of $\boldsymbol{w}(t)$ can be approximated by the finite difference

$$\mathrm{d}\boldsymbol{w}/\mathrm{d}t \approx (\boldsymbol{w}_{k+1} - \boldsymbol{w}_k)/\gamma \,. \tag{3}$$

Substituting into (2) yields the algorithm $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma \boldsymbol{h}(\boldsymbol{w}_k)$. Further replacing $\boldsymbol{h}(\boldsymbol{w})$ by a stochastic oracle leads to the *stochastic approximation* recursion.

SA Recursion: let $\boldsymbol{w}_0 \in \mathbb{R}^d$ be the initialization,

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma_{k+1} \boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}), \; k \in \mathbb{N} \,, \tag{4}$$

where $\gamma_{k+1} > 0$ is the step size, $\boldsymbol{X}_{k+1}$ is a X-valued random variable, and $\boldsymbol{H} : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$ is a vector field that describes stochastic oracles of $\boldsymbol{h}(\cdot)$.

Before delving further into the SA scheme (4), we remark that alternatives to (2) can be used to derive other (stochastic) algorithms for solving (1). For instance, the second order ODE

$$\mathrm{d}^2\boldsymbol{w}/\mathrm{d}t^2 + (3/t)\,\mathrm{d}\boldsymbol{w}/\mathrm{d}t = \boldsymbol{h}(\boldsymbol{w}(t)) \tag{5}$$

with the Taylor approximation formulas

$$\frac{\boldsymbol{w}_{k+1} - \boldsymbol{w}_k}{\gamma} \approx \frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} + \frac{\gamma}{2}\frac{\mathrm{d}^2\boldsymbol{w}}{\mathrm{d}t^2}, \quad \frac{\boldsymbol{w}_k - \boldsymbol{w}_{k-1}}{\gamma} \approx \frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} - \frac{\gamma}{2}\frac{\mathrm{d}^2\boldsymbol{w}}{\mathrm{d}t^2}$$

lead to the Nesterov's accelerated method [44], and admit the same fixed points as (2). Notice that an active area of research is to understand different momentum methods from an ODE perspective [45]. However, in the stochastic setting, there are limited results even in the stochastic gradient case, see [46].

Our focus is on the behavior of the SA scheme (4) derived from (2). To give some insights, it is useful to write the stochastic oracle as $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) = \boldsymbol{h}(\boldsymbol{w}_k) + \mathbf{u}_{k+1}$ where $\mathbf{u}_{k+1}$ is a perturbation vector that distorts the mean field update direction $\boldsymbol{h}(\boldsymbol{w}_k)$. In the simplest setting, the conditional expectation given the past history of the algorithm evaluates to $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] = \boldsymbol{h}(\boldsymbol{w}_k)$ for any $\boldsymbol{w}_k \in \mathbb{R}^d$; the past history $\mathcal{F}_k$ is the filtration $\sigma(\boldsymbol{w}_0, \boldsymbol{X}_\ell, \ell \leq k)$ up to the k-th iteration. For this case, the conditional expectation

$$\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{w}_{k+1}\right] = \boldsymbol{w}_k + \gamma_{k+1}\boldsymbol{h}(\boldsymbol{w}_k) \tag{6}$$

coincides with the deterministic algorithm obtained by discretizing (2) with (3). As such, intuitively (4) may have similar behavior as the ODE flow in (2). Furthermore, $\{\mathbf{u}_{k+1},\ k+1 \in \mathbb{N}\}$ is a martingale difference sequence with respect to (*w.r.t.*) the filtration $\{\mathcal{F}_k,\ k \in \mathbb{N}\}$. Then, (4) is a Cauchy-Euler approximation for solving (2) with the stepsize sequence $\{\gamma_k,\ k \in \mathbb{N}\}$.

Eq. (4) gives a prototype *stochastic algorithm* for tackling many ML and SP problems. A general algorithm design procedure is to find a desired $\boldsymbol{h}(\cdot)$ and embed the problem at hand into (1). Subsequently, we design the stochastic field $\boldsymbol{H}(\cdot)$ to approximate $\boldsymbol{h}(\cdot)$ and apply the SA recursion (4). Note that the design of the stochastic field shall also respect practical constraints such as limited computation complexity, hardware limitation on arithmetic, and the availability of stochastic samples, to list a few.

The stepsize sequence $(\gamma_k)_{k\geq 1}$ plays a critical role on the behavior of the SA scheme, and is thus one of the most important hyper-parameters. One motivation of the theoretical analysis is to propose theoretically grounded rules to tune this sequence, and possibly to ensure that a given choice results in a (worst case) optimal behavior. In the SA case, from a high-level standpoint, there exists a tradeoff between the average process $\mathbf{w}_{k+1} = \mathbf{w}_k + \gamma_{k+1}\boldsymbol{h}(\mathbf{w}_k)$, which necessitates sufficiently large steps to converge towards a critical point (e.g., $\sum_{k\geq 1}\gamma_k = \infty$), and the bias and randomness of the oracle $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$, which typically requires limiting the magnitude of the learning rate (e.g., $\sum_{k\geq 1}\gamma_k^2 < \infty$).

One example is the stochastic optimization problem $\min_{\boldsymbol{w}\in\mathbb{R}^d} F(\boldsymbol{w})$ where $F(\boldsymbol{w}) := \mathbb{E}\left[\ell(\boldsymbol{w}, \boldsymbol{X})\right]$. We can set the mean field $\boldsymbol{h}(\boldsymbol{w}) := -\nabla F(\boldsymbol{w})$ as the gradient. Under regularity conditions, the stochastic oracle $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ satisfying $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] = \boldsymbol{h}(\boldsymbol{w}_k)$ is a *stochastic gradient* defined by $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := -\mathsf{D}_{10}\ell(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ at the $(k+1)$th iteration, and $\boldsymbol{X}_{k+1}$ is an i.i.d. random sample following the same law of $\boldsymbol{X}$ in $\mathbb{E}\left[F(\boldsymbol{w}, \boldsymbol{X})\right]$. In this case, the SA recursion (4) yields the popular stochastic gradient (SG) algorithm [12]. However, the SA recursion (4) is not limited to the SG algorithms: it can cover more general scenarios, e.g., the function $\boldsymbol{h}(\boldsymbol{w}_k)$ is not necessarily the gradient for any objective function in an optimization problem, or the expectation $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right]$ is not a gradient map.

TABLE I: Summary of notations used in the general analysis of SA, in Sections II-A, III-A and IV-A.

Notation	Object	Def. in
		In Section II-A
h	Mean field	eq. (1)
d	Problem dimensionality	eq. (1)
$(\boldsymbol{w}_k)$	Seq. of iterates	eq. (4)
(γ_k)	Seq. of step-size	eq. (4)
$(\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}))$	Stochastic oracle on the field	eq. (4)
$\mathbf{u}_{k+1}$	Noise perturbation	
$\boldsymbol{X}_{k+1}$	X-valued random variable	eq. (4)
$\mathcal{F}_k = \sigma(\boldsymbol{w}_0, (\boldsymbol{X}_\ell)_{\ell\leq k})$	filtration adap. to $(\boldsymbol{w}_k)$	eq. (4)
		In Section III-A
$\mathrm{W} : \mathbb{R}^d \to \mathbb{R}_+$	non-negative Borel function	
$c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1} \in \mathbb{R}_+$	Constants controlling the mean field $\boldsymbol{h}$	**H** 1
$\tau_{0,k}, \tau_{1,k} \in \mathbb{R}_+$	Seq. of constants controlling the expected oracle field	**H** 1
$\sigma_0^2, \sigma_1^2 \in \mathbb{R}_+$	Constants controlling the variance of the oracle field	**H** 1
V	smooth Lyapunov function	**H** 2
L_{V}	smoothness of V	**H** 2
ϱ	link between W and V	**H** 2
c_{V}	upper bound on $\nabla\mathrm{V}$ w.r.t. W	eq. (46)
Λ_{V}	set of points where $\langle\nabla\mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w})\rangle = 0$	eq. (47)
$\mathrm{EQ}(\boldsymbol{h})$	set of equilibrium points of the vector field $\boldsymbol{h}$	eq. (48)
$\varepsilon > 0$	target precision	
R	(random) stopping time	
T	total number of iterations	
R_T	stopping rule strategy after T steps	
		Section IV-A
$(\mathrm{b}_\ell, \eta_\ell)_{\ell\in\{0,1\}}, \gamma_{\max}, \omega_k$	Constants function of $(\sigma_\ell^2, \tau_\ell, c_{\boldsymbol{h},\ell})_{\ell\in\{0,1\}}$, $L_{\mathrm{V}}, c_{\mathrm{V}}$	eqs. (73) to (76)
$\bar{\boldsymbol{w}}_T$	weighted average of the parameters	Remark 3
$\mathscr{V}$	Initial condition term	eq. (83)
B	Non vanishing bias	eq. (83)
λ_k, b_k	Time dependent constants in Theorem 2	eqs. (91), (92)
$\Lambda_{j:k}$	Shortcut for $\prod_{i=j}^k \lambda_i$	
$\beta \in (0, 1]$	Step-size decrease rate	(96)

For readability, Table I aggregates all notations that are used for the generic analysis of the SA scheme, specifically in Sections II-A, III-A, IV-A, V and VI.

B. Motivating Examples

1) Stochastic Gradient Descent: As a warm-up, we begin our exposition on SA schemes with stochastic gradient algorithms, the simplest yet most popular setting. Stochastic gradient algorithm is the workhorse of modern machine learning and data-driven optimization [7], [10], [47]. Much of the success is due to its broad applicability – the stochastic gradient algorithm generally works for any problem for which there is an unbiased gradient estimator. For convex problems, a long line of work [11], [48]–[54] sheds lights on convergence properties of SG, and they are by now well-understood; while non-convex problems are discussed in [10], [55].

To describe the general setting, the mean field $\boldsymbol{h}(\boldsymbol{w})$ of stochastic gradient algorithm as an SA scheme corresponds to the gradient of an objective function that we aim at minimizing. We consider a differentiable objective function $F : \mathbb{R}^d \to \mathbb{R}$. A necessary condition on a model $\boldsymbol{w}$ to be a minimizer of F is to satisfy $\nabla F(\boldsymbol{w}) = 0$. We consider $\boldsymbol{h}(\boldsymbol{w}) := -\nabla F(\boldsymbol{w})$ and look for points $\boldsymbol{w}$ such that $\boldsymbol{h}(\boldsymbol{w}) = 0$.

We highlight two fundamental situations in the case of discriminative learning.

a) *Expected Risk Minimization (ERM) for streaming data.* The function F is the expected loss ℓ on an observation $\boldsymbol{X}$ (with distribution ρ) of a model $\boldsymbol{w}$: $F(\boldsymbol{w}) := \mathbb{E}[\ell(\boldsymbol{w}, \boldsymbol{X})]$. At iteration $k+1$, a new observation $\boldsymbol{X}_{k+1} \sim \rho$, independent from the past, is revealed. The random field is

$$\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := -\mathsf{D}_{10}\ell(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}). \tag{7}$$

b) *ERM for Batch Data*. The function F is the empirical loss ℓ over a set of observations $(Z_1, \ldots, Z_n)$ for a model $\boldsymbol{w}$: $F(\boldsymbol{w}) := n^{-1} \sum_{i=1}^{n} \ell(\boldsymbol{w}, Z_i)$. At iteration $k+1$, a random index $\boldsymbol{X}_{k+1} \in \{1, \ldots, n\}$, independent from the past and with uniform distribution on $\{1, \cdots, n\}$, is sampled by the learner. The random field is defined as

$$\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := -\mathsf{D}_{10}\ell(\boldsymbol{w}_k, Z_{\boldsymbol{X}_{k+1}}). \tag{8}$$

Remark 1. *Extension to mini-batch SG [56]: at iteration $k+1$, the learner may receive (resp. sample) a number b (called mini-batch size) of observations (resp. indices). For the streaming case, the random field $\boldsymbol{H}$ is then $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := \mathsf{b}^{-1} \sum_{i=1}^{\mathsf{b}} \mathsf{D}_{10}\ell(\boldsymbol{w}_k, Z_{\mathsf{b}k+i})$ where $\{Z_k, \ k \in \mathbb{N}\}$ is a sequence of i.i.d. observations and $\boldsymbol{X}_{k+1} := (Z_{\mathsf{b}k+1}, \ldots, Z_{\mathsf{b}(k+1)})$. For the batch case, it is $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := \mathsf{b}^{-1} \sum_{i \in \boldsymbol{X}_{k+1}} \mathsf{D}_{10}\ell(\boldsymbol{w}_k, Z_i)$, with $\boldsymbol{X}_{k+1}$ is a subset of size b sampled at random with or without replacement in $\{1, \ldots, n\}$. The choice of b leads to tradeoffs between per-iteration computation cost and overall convergence rate, interested readers are referred to [10, Sec. 4] for details.*

In the case of batch data, it is possible to use a non-uniform distribution to sample indices - see, for example, [57]. Moreover, the objective function has a finite-sum structure, and the problem is often rewritten as follows for simplicity:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w}), \qquad F(\boldsymbol{w}) := \tfrac{1}{n} \textstyle\sum_{i=1}^{n} f_i(\boldsymbol{w}) \ , \tag{9}$$

where for all $i \in \{1, \ldots, n\}$, $f_i : \mathbb{R}^d \to \mathbb{R}$, and $f_i(\boldsymbol{w}) := \ell(\boldsymbol{w}, Z_i)$. The random field in (8) can then be written as

$$\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{X}) = -\textstyle\sum_{i=1}^{n} X^i \nabla f_i(\boldsymbol{w}), \tag{10}$$

where $\boldsymbol{X} := (X^1, \ldots, X^n) \in \{0, 1\}^n$ is an n-dimensional binary sampling vector with $\sum_{i=1}^{n} X^i = 1$. If the law of the sampling vector $\boldsymbol{X}$ is such that $\mathbb{E}[X^i] = 1/n$ for any i, then $\mathbb{E}[\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{X})] = -\nabla F(\boldsymbol{w})$. We observe that SA scheme using the above $\boldsymbol{H}(\cdot)$ yields the classical stochastic gradient algorithm for the finite-sum problem (9):

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \gamma_{k+1} \textstyle\sum_{i=1}^{n} X_{k+1}^i \nabla f_i(\boldsymbol{w}_k) \ . \tag{11}$$

We note that the computational complexity of (11) is independent of n, since $\sum_{i=1}^{n} X_{k+1}^i = 1$ and the learner must compute $\nabla f_i(\boldsymbol{w}_k)$ for the only index i satisfying $X_{k+1}^i = 1$.

2) Compressed Stochastic Approximation: We study compressed SA methods where a compression operator is used in the update scheme. The goal is either to reduce transmission, storage, or computational costs. We focus here on instantiating methods for the SG case. As a first example, we consider the Gauss-Southwell coordinate descent estimator. For high-dimensional problems ($d \gg 1$), coordinate descent methods reduce computational complexity by restricting the update to a subset of the coordinates. We consider the same optimization problem as (9). Let $j_{k+1} \in \{1, \ldots, d\}$ is the chosen coordinate in the k-th iteration, we have

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \gamma_{k+1} \nabla_{j_{k+1}} F(\boldsymbol{w}_k) \, \boldsymbol{e}_{j_{k+1}} \ , \tag{12}$$

where $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ is the canonical basis of $\mathbb{R}^d$ and $\nabla_j F$ is the j-th coordinate of the gradient. The *Gauss-Southwell selection* rule [22] uses:

$$j_{k+1} := \operatorname*{Argmax}_{j \in \{1, \ldots, d\}} \left| \nabla_j F \left(\boldsymbol{w}_k \right) \right| \ . \tag{13}$$

This corresponds to a greedy selection procedure, since at each iteration we select a coordinate with the largest directional derivative. The Gauss-Southwell rule, on the other hand, corresponds to a deterministic algorithm: $\boldsymbol{H}(\boldsymbol{w}_k, \sim) := \nabla_{j_{k+1}} F \left(\boldsymbol{w}_k \right) \boldsymbol{e}_{j_{k+1}}$ with j_{k+1} in (13). A straightforward extension to SG is possible by replacing the $\nabla_j F(\boldsymbol{w}_k)$ in (13) with $\sum_{i=1}^{n} X_{k+1}^i \nabla_j f_i(\boldsymbol{w}_k)$ and construct the stochastic oracle as $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := \sum_{i=1}^{n} X_{k+1}^i \nabla_{j_{k+1}} f_i \left(\boldsymbol{w}_k \right) \boldsymbol{e}_{j_{k+1}}$.

The coordinate descent algorithm (12) is a special case of the general compressed SG methods that aim to reduce the storage and/or transmission cost of SG. To formally discuss the general methods, we introduce the concept of *compression operators*. A compression operator on $\mathbb{R}^d$ is a mapping $\boldsymbol{\mathcal{C}} : \mathbb{R}^d \times \mathsf{U} \to \mathbb{R}^{\mathsf{d}}$, where U is a general state space equipped with a sigma field and distribution μ_U. The operator $\boldsymbol{\mathcal{C}}$ is called (random) compression if, for any $\mathbf{x} \in \mathbb{R}^d$, the cost of storing/transmitting $\boldsymbol{\mathcal{C}}(\mathbf{x}, \mathbf{U})$, with $\mathbf{U} \sim \mu_U$, is almost-surely (or on average) less than the cost of storing/transmitting $\mathbf{x}$ itself.

Remark 2. *Two prevailing strategies have been used and combined to create such compression schemes [58], [59]:*

1) *(random) projection: $\mathbf{x} \in \mathbb{R}^d$ is projected onto a smaller dimensional subspace. E.g., the $\mathrm{Rand}_{\mathrm{h}}$ operator projects $\mathbf{x}$ onto a random space generated by h canonical vectors; and $\mathrm{Top}_{\mathrm{h}}$ operator projects $\mathbf{x}$ onto a deterministic space generated by the h canonical vectors corresponding to the largest values of $\mathbf{x}$. For example, (13) corresponds to using Top_1 compressor;*
2) *(random) quantization, that (randomly) maps each coordinate of $\mathbf{x}$ onto a scalar codebook and transmits the index of the corresponding codeword [60]. E.g., assuming a uniform quantizer converts a floating point value $x \in \mathbb{R}$ to the closest quantized value as:*

$$Q_d(x, \sim) := \mathrm{sign}(x) \, \Delta \left\lfloor \frac{|x|}{\Delta} + \frac{1}{2} \right\rfloor , \tag{14}$$

where Δ denotes the quantization resolution. On the other hand, the quantization function for stochastic rounding is defined as:

$$Q_s(x, U) := \Delta \times \begin{cases} \left\lfloor \frac{x}{\Delta} \right\rfloor + 1 & \text{if } U \leq \frac{x}{\Delta} - \left\lfloor \frac{x}{\Delta} \right\rfloor , \\ \left\lfloor \frac{x}{\Delta} \right\rfloor & \text{otherwise,} \end{cases} \tag{15}$$

where U is uniformly distributed on $[0, 1)$. These scalar quantizers can be extended to the vector case by applying the quantization operation on each coordinate.

Let $\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{X})$ denotes the SA random field. We introduce three classes of compressed SA methods as follows. The first class corresponds to compressing the random field:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma_{k+1} \boldsymbol{\mathcal{C}}(\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}), \mathbf{U}_{k+1}) \ . \tag{16}$$

It includes (12) as a special case. In the literature, these methods have attracted attention with the increasing interest in distributed optimization, *e.g.,* inexact gradient descent methods [61], [62], low-precision coordinate descent methods [63], compressed SG methods [58], [59], [64]–[68], quantized algorithms for wireless sensor networks [69]–[72] and the references therein.

The second class of compressed SA refers to recursion where $\boldsymbol{H}$ is observed at a point (slightly) different from $\boldsymbol{w}_k$ called the *perturbed iterate*:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma_{k+1}\, \boldsymbol{H}(\boldsymbol{\mathcal{C}}(\boldsymbol{w}_k, \mathrm{U}_{k+1}), \boldsymbol{X}_{k+1})\,. \tag{17}$$

In the SG for deep learning, the above setup is known as a Straight-Through Estimator (STE) introduced by [20], which quantizes the model $\boldsymbol{w}$ before computing the gradient oracle; then the SG update is performed using a full precision buffer. In the convex optimization literature, recursion has been studied as perturbed iterations by [73], and also in the randomized- smoothing approach, where the observation point is intentionally perturbed (e.g., by a Gaussian noise) to achieve better regularity [74], [75]. The same approach includes the study of SG with asynchrony (i.e., the field can be measured on an 'old' iterated model) [76], [77] or in distributed systems where the gradient is observed on a local model held only by the local workers [66], [78].

The third class of compressed SA method is the recursion:

$$\boldsymbol{w}_{k+1} = \boldsymbol{\mathcal{C}}(\boldsymbol{w}_k + \gamma_{k+1}\, \boldsymbol{H}(\boldsymbol{w}_k, , \boldsymbol{X}_{k+1}), \mathrm{U}_{k+1})\,. \tag{18}$$

Note that (18) is a special case of (4) with the random field $\widetilde{\boldsymbol{H}}(\boldsymbol{w}_k, \mathrm{U}_{k+1}, \boldsymbol{X}_{k+1})$ given by

$$\frac{1}{\gamma_{k+1}} \left(\boldsymbol{\mathcal{C}}(\boldsymbol{w}_k + \gamma_{k+1}\, \boldsymbol{H}(\boldsymbol{w}_k, , \boldsymbol{X}_{k+1}), \mathrm{U}_{k+1}) - \boldsymbol{w}_k\right). \tag{19}$$

In the SG case, the above setup is known as the low-precision SG, introduced in [79]–[82], which quantizes the model $\boldsymbol{w}$ after computing the gradient oracle.

The use of low-precision arithmetic plays an essential role in SP and ML, where frugal algorithms are often mandatory. To train models with a low-precision representation of the parameters, we can apply the quantization function $\boldsymbol{\mathcal{C}}(\cdot, \mathrm{U})$ to convert entries of the parameter vector $\boldsymbol{w}$ into a quantized/rounded version $\hat{\boldsymbol{w}} = \boldsymbol{\mathcal{C}}(\boldsymbol{w}, \mathrm{U})$. The first application of the STE rule was in the BinaryConnect algorithm of [20] for training neural networks with Boolean weights; in this case, the weights are binary $\{-\Delta, \Delta\}$. STE has been applied to many different settings and improved; see [83]–[85].

3) Stochastic EM algorithms: The Expectation-Maximisation algorithm (EM) proposed by the popular work [86] was used to solve the optimization problem $\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} F(\boldsymbol{\theta})$ when F is defined by an (possibly intractable) integral $F(\boldsymbol{\theta}) := -\log \int_{\widetilde{\mathcal{Z}}} p(z;\boldsymbol{\theta})\tilde{\mu}(\mathrm{d}z)$, where $p(z;\boldsymbol{\theta})$ is a positive function and $\tilde{\mu}$ is a sigma-finite measure on a measurable set $\widetilde{\mathcal{Z}}$; see [87], [88] and references therein. There are numerous applications of EM, including inference of mixture distributions [89], robust inference in the presence of heavy tailed noise [90], Hidden Markov Models [91], [92], factor analysis [93], graphical models with missing data. In the subsequent discussions, we consider the case where both the function $p(z;\boldsymbol{\theta}) = \prod_{i=1}^n p_i(z_i;\boldsymbol{\theta})$ and the dominating measure $\tilde{\mu} = \mu^{\otimes n}$ have a product form that yields

$$F(\boldsymbol{\theta}) := -\tfrac{1}{n}\textstyle\sum_{i=1}^n \log g_i(\boldsymbol{\theta}), \tag{20}$$

where

$$g_i(\boldsymbol{\theta}) := \textstyle\int_{\mathcal{Z}} p_i(z_i;\boldsymbol{\theta})\, \mu(\mathrm{d}z_i), \tag{21}$$

and $p_i(z_i;\boldsymbol{\theta})$ is positive for all $i \in \mathbb{N}$ and $(z_i, \boldsymbol{\theta}) \in \mathcal{Z} \times \mathbb{R}^d$. Such an optimization problem is motivated by minimizing the Kullback-Leibler divergence computed along the examples indexed by $i \in \mathbb{N}$. As expressed by (20)-(21), in EM the divergence/loss F is not explicit and is given by an integral over a *latent variable* $z = (z_1, \cdots, z_n)$.

A popular application of EM is the computation of Maximum Likelihood estimator. In this case, $g_i(\boldsymbol{\theta})$ is the log-likelihood function of an observation Y_i in a latent variable model (see, e.g., [86]–[88], [94]–[96]). The positive quantity $p_i(z_i;\boldsymbol{\theta})$ is the joint probability of the observation Y_i and the latent variable z_i for a given value of the parameter $\boldsymbol{\theta}$; it is a shorthand notation for $p_{Y_i}(z_i;\boldsymbol{\theta})$. Subsequently, $F(\boldsymbol{\theta})$ corresponds to the case where the observations are independent and the statistical analysis is based on a given set of n examples that are not necessarily identically distributed.

EM is a *Majorize-Minimization* algorithm (see, e.g., [23, chapter 8]) that handles the minimization of (20) by iterating between an *Expectation step* (E-step) and a *Minimization step* (M-step). Given the current iterate $\boldsymbol{\theta}_k$, the E-step defines a surrogate function $\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$ such that: $F(\boldsymbol{\theta}_k) = \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k)$ and $F(\boldsymbol{\theta}) \le \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. The M-step updates the parameter by selecting a minimizer

$$\boldsymbol{\theta}_{k+1} := \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d}\, \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})\,, \tag{22}$$

which is assumed to be unique for simplicity. Under regularity conditions, since $\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\cdot)$ majorizes $F(\cdot)$, $\boldsymbol{\theta}_{k+1}$ is a stationary point of the difference $\boldsymbol{\theta} \mapsto \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}) - F(\boldsymbol{\theta})$. This yields

$$\nabla F(\boldsymbol{\theta}_k) = \nabla \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k)\,.$$

As the surrogate $\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\cdot)$ is an upper bound on the objective, an improvement on the surrogate translates to an improvement on the objective. This descent property lends EM algorithm with remarkable numerical stability. Since each iteration is defined as a minimization of a function, EM algorithm is invariant under changes of parametrization, which is a significant advantage over first-order (gradient) method. The EM surrogate used by [86] is given up to an additive constant (which depends on $\boldsymbol{\theta}'$) as

$$\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{Z}} \log p_i(z_i;\boldsymbol{\theta})\, \pi_i(z_i;\boldsymbol{\theta}')\, \mu(\mathrm{d}z_i), \tag{23}$$

where $z \mapsto \pi_i(z;\boldsymbol{\theta})$ is the probability density function (p.d.f.) on $\mathcal{Z}$ defined by

$$\pi_i(z;\boldsymbol{\theta}) := p_i(z;\boldsymbol{\theta})/g_i(\boldsymbol{\theta}). \tag{24}$$

In many applications, $\pi_i(z_i;\boldsymbol{\theta})$ is the posterior distribution of the latent variable z_i given the observation $\#i$ when the value

of the parameter is $\boldsymbol{\theta}$. A useful decomposition of this surrogate, which is the key discovery of [86], is given by

$$\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log g_i(\boldsymbol{\theta}) + \mathcal{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) \tag{25}$$

where $\mathcal{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$ is the cross-entropy of the distribution $\pi(z_1^n;\boldsymbol{\theta}) := \prod_{i=1}^n \pi_i(z_i;\boldsymbol{\theta})$ relative to the distribution $\pi(\cdot;\boldsymbol{\theta}')$:

$$\mathcal{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{Z}} \log(\pi_i(z_i;\boldsymbol{\theta}))\, \pi_i(z_i;\boldsymbol{\theta}')\, \mu(\mathrm{d}z_i)\ .$$

The cross-entropy $\boldsymbol{\theta} \mapsto \mathcal{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$ is minimized at $\boldsymbol{\theta}'$ such that $\mathcal{H}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}')$ is the entropy of $\pi(\cdot;\boldsymbol{\theta}')$. Under appropriate regularity conditions, this in particular implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\nabla \mathcal{H}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = 0\ . \tag{26}$$

Moreover, the surrogate decomposition (25) and the inequality $\mathcal{H}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_{k+1}) \geq \mathcal{H}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k)$ imply

$$\begin{aligned} -\tfrac{1}{n}\textstyle\sum_{i=1}^n \log g_i(\boldsymbol{\theta}_{k+1}) + \tfrac{1}{n}\sum_{i=1}^n \log g_i(\boldsymbol{\theta}_k) \\ \leq \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_{k+1}) - \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k) \leq 0\ , \end{aligned}$$

where we used the definition of $\boldsymbol{\theta}_{k+1}$ in the RHS. Hence, any update of the EM algorithm leads to an increase of the function $\boldsymbol{\theta} \mapsto \frac{1}{n}\sum_{i=1}^n \log g_i(\boldsymbol{\theta})$.

The limiting point of the EM algorithms are the fixed point of the EM mapping, *i.e.,* the parameters $\boldsymbol{\theta}_\star$ which satisfy

$$\boldsymbol{\theta}_\star = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_\star}(\boldsymbol{\theta}).$$

Under appropriate regularity conditions (see *e.g.,* [97]), we have from (22) and (25) that

$$\nabla \mathcal{H}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_{k+1}) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla \log g_i(\boldsymbol{\theta}_{k+1}). \tag{27}$$

If $\boldsymbol{\theta}_\star$ is a fixed point of (22), together with (26), we have

$$0 = \nabla \mathcal{H}_{\boldsymbol{\theta}_\star}(\boldsymbol{\theta}_\star) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla \log g_i(\boldsymbol{\theta}_\star)\ , \tag{28}$$

showing that the fixed points of the EM coincide with the roots of the gradient of the objective function (see (20)).

We restrict our attention to the case where p_i belongs to the curved exponential family (see, e.g., [98]). Exponential family models are important special cases as the E-step amounts only to computing a conditional expectation. In particular,

EM 1. *There exist measurable functions $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$, $\psi : \mathbb{R}^d \to \mathbb{R}$, and for all $i \in \{1, \cdots, n\}$, $\mathsf{S}_i : \mathcal{Z} \to \mathbb{R}^d$ such that*

$$\log p_i(z_i;\boldsymbol{\theta}) := \langle \mathsf{S}_i(z_i) \,|\, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle - \psi(\boldsymbol{\theta}).$$

In the terminology used for exponential families, the functions S_i's are called the *sufficient statistics*. In the applications, S_i depends on i through the observation Y_i. Moreover,

EM 2. *There exists a measurable function $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$ such that for any $s \in \mathbb{R}^d$*

$$\mathsf{T}(s) := \mathrm{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d}\{\psi(\boldsymbol{\theta}) - \langle s \,|\, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle\}.$$

In most cases, the optimization problem defined for T is strongly convex, and in some cases, it can be solved in closed form. Under **EM** 1, the $\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}'}$ function writes

$$\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) := \psi(\boldsymbol{\theta}) - \langle \bar{\mathsf{s}}(\boldsymbol{\theta}') \,|\, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle,$$

where $\bar{\mathsf{s}}(\boldsymbol{\theta}')$ is the mean value of the expectations of the S_i functions under the p.d.f. π_i's:

$$\bar{\mathsf{s}}(\boldsymbol{\theta}') := \frac{1}{n}\sum_{i=1}^n \bar{\mathsf{s}}_i(\boldsymbol{\theta}'), \tag{29}$$

$$\text{where}\quad \bar{\mathsf{s}}_i(\boldsymbol{\theta}') := \int_{\mathcal{Z}} \mathsf{S}_i(z_i)\, \pi_i(z_i;\boldsymbol{\theta}')\mu(\mathrm{d}z_i). \tag{30}$$

At each iteration of the EM algorithm, the computation of the surrogate function $\boldsymbol{\theta} \to \mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$ boils down to the computation of the expectation $\bar{\mathsf{s}}(\boldsymbol{\theta}_k)$. **EM** 1 and 2 imply that a step of the EM algorithm is expressed as

$$\boldsymbol{\theta}_{k+1} = \mathsf{T} \circ \bar{\mathsf{s}}(\boldsymbol{\theta}_k),$$

thus showing that the fixed points of the EM mapping are the roots of $\boldsymbol{\theta} \mapsto \mathsf{T} \circ \bar{\mathsf{s}}(\boldsymbol{\theta}) - \boldsymbol{\theta}$.

Note that $\bar{\mathsf{s}}(\boldsymbol{\theta}_k)$ might be seen as a double expectation: the *inner* expectation amounts to evaluating $\bar{\mathsf{s}}_i(\boldsymbol{\theta}')$, and the *outer* integral is an average over the n functions $\bar{\mathsf{s}}_i$. In large-scale learning, the outer integral is intractable or has prohibitive computational cost; it may also be the case that the inner integrals are not explicit, e.g., when π_i is known except for a normalizing constant, and its expression or the geometry of $\mathcal{Z}$ is complicated. The *Stochastic EM* algorithms were developed to avoid a full scan of the n functions at each iteration and to allow a stochastic approximation of the inner integrals by Monte Carlo sampling. The stochastic EM algorithms described in [25], [99]–[101] address the intractability of inner expectation; the algorithms in [26], [31]–[33], [102]–[104] address the intractability of the outer expectation; [38] addresses both intractabilities.

Many if not all stochastic EM algorithms are instances of SA. The key ingredient is the following result (see, for example, [25, section 7]): if $\boldsymbol{\theta}_\infty$ is a root of $\boldsymbol{\theta} \mapsto \mathsf{T} \circ \bar{\mathsf{s}}(\boldsymbol{\theta}) - \boldsymbol{\theta}$ on $\mathbb{R}^d$, then $\boldsymbol{w}_\infty := \bar{\mathsf{s}}(\boldsymbol{\theta}_\infty)$ is a root of $\boldsymbol{w} \mapsto \bar{\mathsf{s}} \circ \mathsf{T}(\boldsymbol{w}) - \boldsymbol{w}$ on $\mathbb{R}^d$; *(ii)* if $\boldsymbol{w}_\infty$ is a root of $\boldsymbol{w} \mapsto \bar{\mathsf{s}} \circ \mathsf{T}(\boldsymbol{w}) - \boldsymbol{w}$ on $\mathbb{R}^d$, then $\boldsymbol{\theta}_\infty := \mathsf{T}(\boldsymbol{w}_\infty)$ is a root of $\boldsymbol{\theta} \mapsto \mathsf{T} \circ \bar{\mathsf{s}}(\boldsymbol{\theta}) - \boldsymbol{\theta}$ on $\mathbb{R}^d$. Consequently, EM can be executed in the s-space by running a SA algorithm to solve the root-finding problem

$$\boldsymbol{w} \in \mathbb{R}^d, \qquad \bar{\mathsf{s}} \circ \mathsf{T}(\boldsymbol{w}) - \boldsymbol{w} = 0. \tag{31}$$

Stochastic EM algorithms define a sequence $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ by the iteration $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma_{k+1}\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ where $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ is a random oracle of $\boldsymbol{h}(\boldsymbol{w}_k) := \bar{\mathsf{s}} \circ \mathsf{T}(\boldsymbol{w}_k) - \boldsymbol{w}_k$, the mean field $\boldsymbol{h}$ evaluated at the current iterate $\boldsymbol{w}_k$; and $\{\gamma_k,\ k \in \mathbb{N}\}$ is a deterministic stepsize sequence. The close links between these versions of EM in the s-space and mirror descent algorithms are highlighted in the recent work [105].

To develop a stochastic EM algorithm from (4), we note the randomness $\boldsymbol{X}_{k+1}$ defines a stochastic approximation of the inner and/or outer expectations in the map $\bar{\mathsf{s}}$, see (29). Given a SA sequence $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ converging to a solution $\boldsymbol{w}_\star$ of the fixed point problem (31), the sequence $\boldsymbol{\theta}_k := \mathsf{T}(\boldsymbol{w}_k)$ converges to $\boldsymbol{\theta}_\star$ which is, thanks to (28), a stationary point of the function F, *i.e.,* $\nabla F(\boldsymbol{\theta}_\star) = 0$. Of course, formulating precisely this technical result requires assumptions on the regularity of the model. Among the many stochastic versions of EM in the literature (see references above), let us make the stochastic mean field $\boldsymbol{H}$ explicit for two of them.

a) Mini-batch EM: This algorithm is an adaptation to the finite-sum context of the *Online EM* [31], which is designed to process a data stream. Mini-batch EM avoids computing the exact outer expectation in (29) at each iteration of EM; it replaces the sum over n terms with a sum over a mini-batch chosen at random. The stochastic oracle is given by

$$\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := \frac{1}{\mathsf{b}_{\mathrm{EM}}} \sum\nolimits_{i \in \boldsymbol{X}_{k+1}} \bar{\mathsf{s}}_i \left(\mathsf{T}(\boldsymbol{w}_k)\right) - \boldsymbol{w}_k, \quad (32)$$

where $\boldsymbol{X}_{k+1}$ is a set of size b_{EM}, collecting indices picked at random with or without replacement in $\{1, \cdots, n\}$. See e.g. [106]–[108] for an application.

b) Stochastic Approximation EM (SAEM): This algorithm, proposed by [25], deals with the case the inner expectations in (29) are intractable and must be approximated by Monte Carlo, while the outer expectations in (29) can be computed with reasonable computational effort; see *e.g.,* [109]–[125] for applications. The stochastic oracle is given by

$$\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\mathsf{m}} \rho_i^j(\boldsymbol{w}_k)\, \mathsf{S}_i(Z_{i,k+1}^j) - \boldsymbol{w}_k \quad (33)$$

where $\boldsymbol{X}_{k+1} := (Z_{i,k+1}^j, 1 \le i \le n, 1 \le j \le \mathsf{m})$ collect the m outputs of n distinct Monte Carlo samplers designed to approximate the distributions $\pi_i(z_i; \mathsf{T}(\boldsymbol{w}_k))$, for $i = 1, \ldots, n$.

If i.i.d. sampling is possible from the p.d.f. $\pi_i(z_i; \mathsf{T}(\boldsymbol{w}_k))$, then $\rho_i^j(\boldsymbol{w}_k) := 1/\mathsf{m}$. If sampling according to the conditional distribution $\pi_i(z_i; \mathsf{T}(\boldsymbol{w}_k)))$ is not possible, more sophisticated sampling techniques must be used. For illustration, we consider below the importance sampling procedure, a method of using independent samples from a proposal distribution $\tilde{\pi}_i(z_i; \mathsf{T}(\boldsymbol{w}_k))$ to approximate the expectation with respect to the target distribution $\pi_i(z_i; \mathsf{T}(\boldsymbol{w}_k))$. The importance sampling estimator approximates the target distribution by a random probability measure using weighted samples that are generated from the proposal; (see e.g. [126, Chapter V] and [127]). More precisely, the self-normalized Importance Sampling estimator works as follows. We first sample independently $Z_{i,k+1}^1, \ldots, Z_{i,k+1}^{\mathsf{m}}$ from the proposal $\tilde{\pi}_i(z_i; \mathsf{T}(\boldsymbol{w}_k))$ and define the normalized importance weights

$$\rho_i^j(\boldsymbol{w}_k) := \frac{p_i(Z_{i,k+1}^j; \mathsf{T}(\boldsymbol{w}_k))}{\tilde{\pi}_i(Z_{i,k+1}^j; \mathsf{T}(\boldsymbol{w}_k))} \left(\sum_{\ell=1}^{\mathsf{m}} \frac{p_i(Z_{i,k+1}^\ell; \mathsf{T}(\boldsymbol{w}_k))}{\tilde{\pi}_i(Z_{i,k+1}^\ell; \mathsf{T}(\boldsymbol{w}_k))} \right)^{-1}. \quad (34)$$

Instead of IS, another option is to use Markov Chain Monte Carlo (MCMC) samplers targeting the p.d.f. $\pi_i(z_i; \mathsf{T}(\boldsymbol{w}_k))$; see for example [25], [26], [101], [104], [109].

4) TD Learning: Many tasks in reinforcement learning such as policy evaluation, Q-learning [29], etc., can be formulated as root finding problems whose effective solutions are often given as non-stochastic-gradient SA recursions. Below, we select the temporal difference (TD) learning algorithm [28] for policy evaluation to illustrate another aspect of general principle of stochastic algorithm design.

We follow the derivations of [128] in this example. Consider the problem of evaluating the value function of applying a policy π in a Markov Decision Process. The policy π specifies the conditional probability of choosing an action given a certain state, It induces a Markov Reward Process (MRP) given by the tuple $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \lambda)$: $\mathcal{S} = \{s_1, \ldots, s_n\}$ is the state-space (assumed for simplicity to be finite); $\mathcal{P}$ is the $n \times n$ state transition matrix of the probability of transition from a given state to another; $\mathcal{R}$ is the reward function such that $\mathcal{R}(s, s')$ associates a reward with each state transition; $\lambda \in (0, 1)$ is the discount factor. With a slight abuse of notation, we define the expected instantaneous reward from state s, for $s \in \mathcal{S}$,

$$\mathcal{R}(s) := \textstyle\sum_{s' \in \mathcal{S}} \mathcal{P}(s, s') \mathcal{R}(s, s');$$

As a standing assumption, we concentrate on ergodic MRPs:

TD 1. *The non-negative matrix $\mathcal{P}$ is irreducible and has a unique stationary distribution ϖ,* i.e., $\varpi \mathcal{P} = \varpi$.

Note that $\varpi(s) > 0$ for any $s \in \mathcal{S}$ under **TD** 1.

The value function $\mathcal{V}$ of the above MRP is the expected cumulative discounted reward for a given state $s \in \mathcal{S}$, *i.e.,*

$$\mathcal{V}(s) := \mathbb{E}\left[\sum_{k=0}^{\infty} \lambda^k \mathcal{R}(S_k) \,\middle|\, S_0 = s \right] = \sum_{k=0}^{\infty} \lambda^k \mathcal{P}^k \mathcal{R}(s), \quad (35)$$

where the expectation is over the distribution of the Markov chain $\{S_k,\ k \in \mathbb{N}\}$, started at $S_0 = s$, with Markov kernel $\mathcal{P}$. Note that $\mathcal{P}^k \mathcal{R}(s) = \sum_{s' \in \mathcal{S}} \mathcal{P}^k(s, s') \mathcal{R}(s')$. This value function obeys the Bellman equation $\mathrm{B}\,\mathcal{V} = \mathcal{V}$ where the Bellman operator B is defined as

$$[\mathrm{B}\, V](s) := \mathcal{R}(s) + \lambda \textstyle\sum_{s' \in \mathcal{S}} \mathcal{P}(s, s') V(s'),\ \forall\ s \in \mathcal{S}\ , \quad (36)$$

for any function $V : \mathcal{S} \to \mathbb{R}$. Assume bounded reward function, the value function $\mathcal{V}(s)$ is well defined and is the only solution to (36) [29], [129]. In most applications, the cardinality n of the state space $\mathcal{S}$ is large. It is often advocated in such case to resort to parametric approximations, for example by using a linear function approximation or deep neural networks [130], [131]. For simplicity, we focus on linear function approximation, where $\mathcal{V}(s) \approx \mathcal{V}_{\boldsymbol{w}}(s) := \boldsymbol{\phi}(s)^\top \boldsymbol{w}$; $\boldsymbol{\phi}(s) \in \mathbb{R}^d$ is called the *feature vector* for the state $s \in \mathcal{S}$, and $\boldsymbol{w} \in \mathbb{R}^d$ is a parameter vector to be estimated. Without loss of generality, we assume that

TD 2. *For all $s, s' \in \mathcal{S}$, $|\mathcal{R}(s, s')| \le 1$, $\max_{s \in \mathcal{S}} \|\boldsymbol{\phi}(s)\| \le 1$.*

Since the state space is finite, the value function $\mathcal{V}_{\boldsymbol{w}}$ can be represented as a vector in $\mathbb{R}^n$, whose i-th coordinate is $\mathcal{V}_{\boldsymbol{w}}(s_i)$. This vector can be written compactly as

$$\mathcal{V}_{\boldsymbol{w}} = \boldsymbol{\Phi} \boldsymbol{w}, \quad \boldsymbol{\Phi} := [\boldsymbol{\phi}(s_1), \ldots, \boldsymbol{\phi}(s_n)]^\top, \quad (37)$$

such that $\boldsymbol{\Phi}$ is the $n \times d$ feature matrix. The linear function approximation restricts the set of admissible value function $\mathcal{V}_{\boldsymbol{w}}$ to $\mathrm{span}(\boldsymbol{\Phi})$, the subset of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{\Phi}$. As a result, the Bellman equation $\mathrm{B}\,\mathcal{V} = \mathcal{V}$ may no longer be satisfied by $\mathcal{V}_{\boldsymbol{w}}$ for any $\boldsymbol{w} \in \mathbb{R}^d$.

Among the many TD learning algorithms (see e.g. [132, Section B]), we consider in this paper the so-called TD(0)-learning algorithm. TD(0) is an SA scheme: under the on-policy setting where the data is generated from the MRP induced by π, the algorithm starts with an initial estimate $\boldsymbol{w}_0$ and at iteration $(k+1)$, it gets a new observation

$\boldsymbol{X}_{k+1} = (S_{k+1}, S'_{k+1})$, and computes the next iterate by $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma_{k+1}\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ where

$$\boldsymbol{H}(\boldsymbol{w}, (s, s')) := \left(\mathcal{R}(s, s') + \lambda\boldsymbol{\phi}(s')^\top \boldsymbol{w} - \boldsymbol{\phi}(s)^\top \boldsymbol{w}\right)\boldsymbol{\phi}(s) \,. \tag{38}$$

We make a simplifying assumption about $\boldsymbol{X}_{k+1}$:

TD 3. *The sequence $\{S_k,\ k \in \mathbb{N}\}$ is sampled independently from the stationary distribution ϖ and, for each k, S'_k is sampled from $\mathcal{P}(S_k, \cdot)$.*

This assumption is classical in reinforcement learning and is suitable for algorithms that use a replay buffer [133].

Under mild conditions, the expected value of S'_{k+1} conditionally to $\{S_{k+1} = s\}$ of $\mathcal{R}(S_{k+1}, S'_{k+1}) + \lambda\boldsymbol{\phi}(S'_{k+1})^\top \boldsymbol{w} - \boldsymbol{\phi}(S_{k+1})^\top \boldsymbol{w}$ evaluate to $[\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}}](s) - \mathcal{V}_{\boldsymbol{w}}(s)$ which gives the *temporal difference error* for the Bellman equation with the estimate $\boldsymbol{w}$. For any function $g : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, denote

$$\mathbb{E}_\varpi[g(S_0, S'_0)] := \textstyle\sum_{s,s' \in \mathcal{S}} \varpi(s)\mathcal{P}(s, s')g(s, s') \,. \tag{39}$$

The mean field function of the TD(0) algorithm is given by

$$\begin{aligned}\boldsymbol{h}(\boldsymbol{w}) &:= \mathbb{E}_\varpi[\boldsymbol{\phi}(S_0)\mathcal{R}(S_0, S'_0)] \\ &\quad + \mathbb{E}_\varpi[\boldsymbol{\phi}(S_0)\{\lambda\boldsymbol{\phi}(S'_0) - \boldsymbol{\phi}(S_0)\}^\top]\,\boldsymbol{w} \\ &= \boldsymbol{\Phi}^\top \mathbf{D}_\varpi(\mathrm{B}\,\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{\Phi}\boldsymbol{w})\end{aligned} \tag{40}$$

where we have set the diagonal matrix

$$\mathbf{D}_\varpi := \mathrm{diag}(\varpi(s_1), \ldots, \varpi(s_n)). \tag{41}$$

If $\boldsymbol{w}_\star$ is a root of $\boldsymbol{h}(\boldsymbol{w}) = 0$, then for all $\boldsymbol{w}' \in \mathbb{R}^d$,

$$\langle \boldsymbol{\Phi}\boldsymbol{w}' \mid \mathrm{B}\,\boldsymbol{\Phi}\boldsymbol{w}_\star - \boldsymbol{\Phi}\boldsymbol{w}_\star \rangle_{\mathbf{D}_\varpi} = 0 \tag{42}$$

showing that the Bellman error $\mathrm{B}\,\boldsymbol{\Phi}\boldsymbol{w}_\star - \boldsymbol{\Phi}\boldsymbol{w}_\star$ is orthogonal to the linear subspace spanned by the column of the feature matrix $\boldsymbol{\Phi}$ in the scalar product $\langle \cdot \mid \cdot \rangle_{\mathbf{D}_\varpi}$. In light of (42), we may now characterize the root $\boldsymbol{w}_\star$ to $\boldsymbol{h}(\boldsymbol{w}) = \mathbf{0}$ using the projected Bellman equation

$$\mathcal{V}_{\boldsymbol{w}} = \mathrm{Prj}_\varpi\,\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}} \,; \tag{43}$$

Prj_ϖ is the projection operator onto $\mathrm{span}(\boldsymbol{\Phi})$ *w.r.t.* $\|\cdot\|_{\mathbf{D}_\varpi}$. In Section III-B4, we will show that the equation $\boldsymbol{h}(\boldsymbol{w}) = 0$ has a unique root, $\boldsymbol{w}_\star$, which is the fixed point to (43).

Summary. In this section, we introduced the general SA scheme under consideration, and showed that it can be instantiated in several major applications of ML and SP. We summarize in Table II the algorithms that we introduce.

III. Assumptions

In this section, we first introduce in Section III-A the main assumptions on the mean field $\boldsymbol{h}$ and the random oracle $\boldsymbol{H}$ under which theoretical results will be derived in Sections IV to VI. We then establish conditions under which those assumptions are satisfied on the four previously detailed examples, in respectively Sections III-B1 to III-B4.

TABLE II: Summary of algorithms introduced in Section II-B, that are variants of the SA scheme (4).

Setting	Name	Reference
SGD	SGD for ERM	eq. (10)
Compressed SA	Gauss-Southwell (GD with Top_1)	eqs. (12) and (13)
	SA with compressed field	eq. (16)
	STE, SA with quantized iterates	eq. (17)
	Low precision SA	eq. (19)
EM	Mini-batch EM	eq. (32)
	SAEM with independent MC	eq. (33)
	SAEM with importance sampling MC	eqs. (33) and (34)
TD	TD(0)	eq. (38)

A. Assumptions on the SA scheme

Assume that the root-finding problem (1) is to be solved by the SA algorithm (4) which acquires the mean-field $\boldsymbol{h}(\boldsymbol{w})$ via subsequent calls to a *stochastic oracle*. At iteration $(k+1)$, we denote $\boldsymbol{w}_k$ as the current value of the model and the stochastic oracle outputs $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$, where $\{\boldsymbol{X}_k,\ k \in \mathbb{N}\}$ are random variables taking values in $\mathsf{X} \subseteq \mathbb{R}^\ell$. We denote by $\mathcal{F}_k := \sigma(\boldsymbol{w}_0, \boldsymbol{X}_\ell, 1 \le \ell \le k)$ the sigma-algebra generated by the random variables $\{\boldsymbol{X}_\ell\}_{\ell=1}^k$ and the initial model $\boldsymbol{w}_0$.

We shall consider a set of assumptions for the Borel functions $\boldsymbol{H} : \mathbb{R}^d \times \mathsf{X} \to \mathbb{R}^d$ and $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^d$ in relation to the SA recursion. This gives rise to the first ingredient of our analysis framework that considers a non-negative Borel function $\mathrm{W} : \mathbb{R}^d \to \mathbb{R}_+$ which controls the growth to infinity of the variance of the stochastic oracle and its bias. We use the same function in **H** 2, where it measures the coercivity of the mean field.

H 1. a) *For all $k \ge 0$, $\mathbb{E}[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\|^2] < \infty$.*

b) *There exist $c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1} \in \mathbb{R}_+$ such that for all $\boldsymbol{w} \in \mathbb{R}^d$*

$$\|\boldsymbol{h}(\boldsymbol{w})\|^2 \le c_{\boldsymbol{h},0} + c_{\boldsymbol{h},1}\mathrm{W}(\boldsymbol{w}).$$

c) *For any $k \ge 0$, there exist $\tau_{0,k}, \tau_{1,k} \in \mathbb{R}_+$ such that, a.s.,*

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 \le \tau_{0,k} + \tau_{1,k}\mathrm{W}(\boldsymbol{w}_k). \tag{44}$$

d) *There exist $\sigma_0^2, \sigma_1^2 \in \mathbb{R}_+$ such that for any $k \ge 0$, a.s.,*

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right]\|^2\right] \\ \le \sigma_0^2 + \sigma_1^2\mathrm{W}(\boldsymbol{w}_k).\end{aligned} \tag{45}$$

Under these assumptions, the random oracle $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ may be a biased estimator of the mean-field $\boldsymbol{h}(\boldsymbol{w}_k)$, with $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] - \boldsymbol{h}(\boldsymbol{w}_k)$ being a possibly time-varying conditional bias. A special case often considered in the literature pertains to unbiased SA where $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] = \boldsymbol{h}(\boldsymbol{w}_k)$. In this case, $\tau_{0,k} = \tau_{1,k} := 0$ for all $k \in \mathbb{N}$ in **H** 1-*c*).

Definition 1 (Unbiased stochastic oracle (USO))**.** *The stochastic oracle in SA scheme is said to be* unbiased *if $\tau_{0,k} = \tau_{1,k} = 0$ for any $k \in \mathbb{N}$.*

In USO, the sequence $\{\boldsymbol{u}_k,\ k \in \mathbb{N}\}$ where $\boldsymbol{u}_{k+1} := \boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) - \boldsymbol{h}(\boldsymbol{w}_k)$, is a martingale increment sequence: $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{u}_{k+1}\right] = 0$, $\mathbb{P}$-a.s.

An example of USO is the case where the mean field $\boldsymbol{h}$ has a finite sum structure: $\boldsymbol{h} = n^{-1}\sum_{i=1}^n \boldsymbol{h}_i$. If $\boldsymbol{X}$ is a uniform

random variable on $\{1,\dots,n\}$, then $\boldsymbol{H}(\boldsymbol{w},\boldsymbol{X}) := \boldsymbol{h}_{\boldsymbol{X}}(\boldsymbol{w})$ is an unbiased stochastic oracle. A generalization of this example is the case where $\boldsymbol{h}$ is defined as an expectation $\boldsymbol{h}(\boldsymbol{w}) := \int S(z,\boldsymbol{w})\,\pi(\mathrm{d}z;\boldsymbol{w})$ with respect to a distribution π that may depend on the current value of the parameter $\boldsymbol{w}$; then $\boldsymbol{H}(\boldsymbol{w},\boldsymbol{X}) := N^{-1}\sum_{i=1}^{N} S(Z_i,\boldsymbol{w})$, where $\boldsymbol{X} := (Z_1,\dots,Z_N)$ ares i.i.d. samples from $\pi(\cdot;\boldsymbol{w})$, induces an unbiased stochastic oracle. If instead a self-normalized importance sampling is used (see *e.g.,* (34)), then the oracle is biased. In this case, the bias is inversely proportional to the number of Monte Carlo samples used in the importance sampling estimate; see *e.g.,* [127].

It is worth noting that in the *Robbins-Monro* setting - in reference to [1] - the sequence $\{\boldsymbol{X}_k,\ k\in\mathbb{N}\}$ is assumed to be i.i.d. In comparison, **H** 1-*c)* and **H** 1-*d)* used here are slightly weaker since we do not assume that $\{\boldsymbol{X}_k,\ k\in\mathbb{N}\}$ are independent nor that they have the same distribution. However, this excludes more subtle dependency structures and time-dependent distributions: in some situations $\{\boldsymbol{X}_k,\ k\in\mathbb{N}\}$ is a Markov chain (possibly) controlled by $\{\boldsymbol{w}_k,\ k\in\mathbb{N}\}$; considering such dependency structures requires the use of sophisticated probabilistic methods, that go beyond the scope of this survey; see [5], [27], [37], [134]–[138] and the references therein.

H 1-*d)* implies the conditional variance of the random oracle $\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})$ is either bounded ($\sigma_1^2=0$) or does not grow faster than $\mathrm{W}(\boldsymbol{w}_k)$. The case $\sigma_1^2=0$ is called the *bounded variance* case; considered in the analysis of SG by [55].

Our analysis for SA follows that of the Lyapunov function approach. In this setup, we introduce the second ingredient of our analysis framework which considers a *smooth* Lyapunov function V for the flow of the nonlinear ODE $\mathrm{d}\mathbf{w}/\mathrm{d}t = \boldsymbol{h}(\mathbf{w})$. Formally, we have the following set of assumptions:

H 2. *There exists a function* $\mathrm{V}:\mathbb{R}^d\to\mathbb{R}$ *such that,*

a) $\mathrm{V}_\star := \inf_{\boldsymbol{w}\in\mathbb{R}^d}\mathrm{V}(\boldsymbol{w}) > -\infty$.

b) V *is continuously differentiable and* L_{V}*-smooth,* i.e., *there exists* $L_{\mathrm{V}}\in\mathbb{R}_+$ *such that for all* $\boldsymbol{w},\boldsymbol{w}'\in\mathbb{R}^d$, $\|\nabla\mathrm{V}(\boldsymbol{w})-\nabla\mathrm{V}(\boldsymbol{w}')\|\le L_{\mathrm{V}}\|\boldsymbol{w}-\boldsymbol{w}'\|$.

c) *There exists* $\varrho>0$ *such that, for all* $\boldsymbol{w}\in\mathbb{R}^d$, $\langle\nabla V(\boldsymbol{w})\,|\,\boldsymbol{h}(\boldsymbol{w})\rangle \le -\varrho\mathrm{W}(\boldsymbol{w})$.

It may seem a bit complicated to have many different constants that may seem redundant: For example, we could have fixed $\mathrm{W}(\boldsymbol{w}) := -\langle\nabla\mathrm{V}(\boldsymbol{w})\,|\,\boldsymbol{h}(\boldsymbol{w})\rangle$ and choose $\varrho=1$, then change the constants $c_{\boldsymbol{h},1}$, $\tau_{1,k}$, σ_1^2, and c_{V} accordingly. The motivation behind **H** 2-*c)* is to add an additional degree of flexibility that will later facilitate analysis and allow covering settings associated with different choices for W.

We will often use the constant $c_{\mathrm{V}}>0$ which satisfies:

$$\|\nabla\mathrm{V}(\boldsymbol{w})\| \le c_{\mathrm{V}}\sqrt{\mathrm{W}(\boldsymbol{w})},\quad \boldsymbol{w}\in\mathbb{R}^d\,. \tag{46}$$

Note that c_{V} might be $+\infty$ (see e.g. Section III-B1 for some examples). On the set

$$\Lambda_{\mathrm{V}} := \{\boldsymbol{w}\in\mathbb{R}^d, \langle\nabla\mathrm{V}(\boldsymbol{w})\,|\,\boldsymbol{h}(\boldsymbol{w})\rangle = 0\}\,, \tag{47}$$

the function W is identically zero by **H** 2-*c)*: for all $\boldsymbol{w}\in\Lambda_{\mathrm{V}}$, $\mathrm{W}(\boldsymbol{w})=0$. However the condition $\mathrm{W}(\boldsymbol{w})=0$ is not sufficient yet to be able to derive useful information for solving the root finding problem $\boldsymbol{h}(\boldsymbol{w})=\mathbf{0}$. The set of equilibrium points of the vector field $\boldsymbol{h}$ is the set

$$\mathrm{EQ}(\boldsymbol{h}) := \{\boldsymbol{w}\in\mathbb{R}^d, \boldsymbol{h}(\boldsymbol{w})=0\} = \{\boldsymbol{h}=0\}\,. \tag{48}$$

Obviously, one has $\mathrm{EQ}(\boldsymbol{h})\subset\Lambda_{\mathrm{V}}$, but the converse may not hold. As we are trying to approach the equilibrium points, it is sensible to assume that V is a *strict Lyapunov* function, *i.e.,*

$$\mathrm{EQ}(\boldsymbol{h}) = \Lambda_{\mathrm{V}}\,. \tag{49}$$

If this is the case, the function W on $\mathrm{EQ}(\boldsymbol{h})$ is zero. In the case of SG, the Lyapunov function V is the objective function to be minimized F. The mean field $\boldsymbol{h}$ is the negated gradient of the objective $\boldsymbol{h}=-\nabla F$. In this case, $\langle\nabla\mathrm{V}\,|\,\boldsymbol{h}\rangle = -\|\nabla F\|^2$ and thus V is automatically a strict Lyapunov function. For the function W, we typically choose $\mathrm{W}(\boldsymbol{w}) := \|\nabla F(\boldsymbol{w})\|^2 = \|\boldsymbol{h}(\boldsymbol{w})\|^2$, so that $\{\mathrm{W}=0\}=\mathrm{EQ}(\boldsymbol{h})$. See Section III-B.

To obtain meaningful results, the function W should be lower bounded outside any open neighborhood of $\mathrm{EQ}(\boldsymbol{h})$: for any $\delta>0$, it is typically required that

$$\min_{\mathrm{d}(\boldsymbol{w},\{\boldsymbol{h}=0\})\ge\delta}\mathrm{W}(\boldsymbol{w}) =: \epsilon_{\mathrm{W}}(\delta) > 0\,, \tag{50}$$

where $\mathrm{d}(\boldsymbol{w},\mathcal{A})$ is the Euclidean distance of $\boldsymbol{w}$ to the set $\mathcal{A}$. We have chosen not to include this condition in **H** 2 because it is not involved in the proofs of the results presented below, but only in their interpretation. Under (50), if for any $\varepsilon>0$, there exists a stopping time R, possibly random, such that $\mathbb{E}\left[\mathrm{W}(\boldsymbol{w}_R)\right]\le\varepsilon$, then for any $\delta>0$, we have

$$\begin{aligned}\mathbb{P}(\mathrm{d}(\boldsymbol{w}_R,\{\boldsymbol{h}=0\})\ge\delta) &\le \mathbb{P}(\mathrm{W}(\boldsymbol{w}_R)\ge\epsilon_{\mathrm{W}}(\delta))\\ &\le \mathbb{E}[\mathrm{W}(\boldsymbol{w}_R)]/\epsilon_{\mathrm{W}}(\delta) \le \varepsilon/\epsilon_{\mathrm{W}}(\delta).\end{aligned}$$

This upper bound can be set arbitrarily small with a convenient choice of ε. We will prove in Section IV that under **H** 1 and **H** 2, given a total number of iterations T, there exists a stopping rule strategy R_T such that $\mathbb{E}\left[\mathrm{W}(\boldsymbol{w}_{R_T})\right]$ goes to zero when $T\to\infty$. To understand why, we analyze the deterministic sequence $\mathbf{w}_{k+1}=\mathbf{w}_k+\gamma_{k+1}\boldsymbol{h}(\mathbf{w}_k)$. Note that, by **H** 2-*b)*,

$$\begin{aligned}\mathrm{V}(\mathbf{w}_{k+1}) \le \mathrm{V}(\mathbf{w}_k) &+ \gamma_{k+1}\langle\nabla\mathrm{V}(\mathbf{w}_k)\,|\,\boldsymbol{h}(\mathbf{w}_k)\rangle\\ &+ (L_{\mathrm{V}}/2)\gamma_{k+1}^2\|\boldsymbol{h}(\mathbf{w}_{k+1})\|^2\,.\end{aligned}$$

Using that $\langle\nabla\mathrm{V}(\mathbf{w})\,|\,\boldsymbol{h}(\mathbf{w})\rangle \le -\varrho\mathrm{W}(\mathbf{w})$ and assuming for simplicity $\|\boldsymbol{h}(\mathbf{w}_{k+1})\|^2\le c_{\boldsymbol{h},0}$, we immediately get that

$$\varrho\gamma_{k+1}\mathrm{W}(\mathbf{w}_k) \le \mathrm{V}(\mathbf{w}_k) - \mathrm{V}(\mathbf{w}_{k+1}) + (L_{\mathrm{V}}/2)\gamma_{k+1}^2 c_{\boldsymbol{h},0}\,.$$

For any $T>0$, setting $\gamma_k=\frac{1}{\sqrt{T}}$ for $k\in\{1,\dots,T\}$, we get

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathrm{W}(\mathbf{w}_k) \le \frac{2\{\mathrm{V}(\mathbf{w}_0)-\mathrm{V}_\star\}+L_{\mathrm{V}}c_{\boldsymbol{h},0}}{2\varrho\sqrt{T}}\,.$$

If R_T is a uniform random variable on $\{0,\dots,T-1\}$, then

$$\mathbb{E}[\mathrm{W}(\mathbf{w}_{R_T})] \le \frac{2\{\mathrm{V}(\mathbf{w}_0)-\mathrm{V}_\star\}+L_{\mathrm{V}}c_{\boldsymbol{h},0}}{2\varrho\sqrt{T}}\,.$$

The RHS goes to zero if $T\to\infty$. This discussion is essentially an anticipation of the results to be obtained in Section IV, which introduce the bias and variance of the random oracle.

B. Verifying Assumptions for the Examples

1) Stochastic Gradient Descent: We recall from Section II-B1 the stochastic gradient algorithm (11). Here, the mean field $\boldsymbol{h}(\boldsymbol{w}) = -\nabla F(\boldsymbol{w})$ is the negated gradient of the objective function. For simplicity, we consider the algorithm for the batch case (8), but results and assumption could easily be extended to include the streaming data framework (7). We concentrate on the case where the sampling of the index $\boldsymbol{X}_{k+1}$ in (8) is uniform over $\{1, \dots, n\}$. The oracle $\boldsymbol{H}(\boldsymbol{w}, \boldsymbol{X})$ of the mean field $\boldsymbol{h}(\boldsymbol{w})$ is then unbiased, satisfying **H** 1-*c)* with $\tau_{0,k} = \tau_{1,k} = 0$. Next, we describe conditions under which **H** 1, 2 are valid and derive the required constants.

Smooth (possibly non-convex) objective functions. We instantiate the assumptions by setting the Lyapunov function as $\mathrm{V}(\boldsymbol{w}) := F(\boldsymbol{w})$ and choosing $\mathrm{W}(\boldsymbol{w}) := \|\nabla F(\boldsymbol{w})\|^2$. With these choices, **H** 1-*b)* is satisfied with $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1}) := (0, 1)$, **H** 2-*c)* is satisfied with $\varrho := 1$ since $\langle \nabla \mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w}) \rangle = -\|\nabla F(\boldsymbol{w})\|^2$. Finally, we have $c_{\mathrm{V}} := 1$ in (46). To establish the other conditions, namely **H** 1-*d)* and **H** 2-*a)*, *b)*, we add the following assumption.

SG 1. a) *For any $i \in \{1, \dots, n\}$, the function f_i is differentiable and its gradient is $L_{\nabla f_i}$-Lipschitz,* i.e., *for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, $\|\nabla f_i(\boldsymbol{w}) - \nabla f_i(\boldsymbol{w}')\| \le L_{\nabla f_i} \|\boldsymbol{w} - \boldsymbol{w}'\|$.*
b) *For any $i \in \{1, \dots, n\}$, $M := \max_{i \in \{1,\dots,n\}} \sup_{\boldsymbol{w} \in \mathbb{R}^d} \|\nabla f_i(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\| < \infty$.*

Thus **H** 1-*d)* holds with $\sigma_0^2 := M^2/n$ and $\sigma_1^2 := 0$. Under the above condition, **H** 2-*a)* holds with $\mathrm{V}_\star := \inf_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w})$. **H** 2-*b)* holds with $L_{\mathrm{V}} := n^{-1} \sum_{i=1}^n L_{\nabla f_i}$.

Smooth convex objective functions. When F is convex, we can choose other functions V, W and obtain different constants in **H** 1-2. We consider

CVX 1. *The function F is convex with an optimal solution* i.e., $\operatorname{Argmin}_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w}) \neq \emptyset$.

Note that a differentiable function is convex on $\mathbb{R}^d$ if and only if for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d \times \mathbb{R}^d$,

$$F(\boldsymbol{w}') \ge F(\boldsymbol{w}) + \langle \nabla F(\boldsymbol{w}) \,|\, \boldsymbol{w}' - \boldsymbol{w} \rangle \;, \tag{51}$$

and equivalently, its gradient is monotone

$$\langle \nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}') \,|\, \boldsymbol{w} - \boldsymbol{w}' \rangle \ge 0 \;; \tag{52}$$

see *e.g.,* Proposition 17.45 and Proposition 17.7 in [139]. Second, under **SG** 1 and **CVX** 1, the mapping ∇F is *co-coercive i.e.,* for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$,

$$\langle \nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}') \,|\, \boldsymbol{w} - \boldsymbol{w}' \rangle \ge (1/L_{\nabla F}) \|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}')\|^2 \;; \tag{53}$$

see *e.g.,* Proposition 17.45 and Corollary 18.14 in [139]. When F is strictly convex, then the inequalities in (52) and (53) are strict (see *e.g.,* Proposition 17.10. in [139]).

Let $\boldsymbol{w}_\star \in \mathbb{R}^d$ be a root of ∇F: $\nabla F(\boldsymbol{w}_\star) = 0$. We get from (53) that $\|\nabla F(\boldsymbol{w})\|^2 \le L_{\nabla F} \langle \nabla F(\boldsymbol{w}) \,|\, \boldsymbol{w} - \boldsymbol{w}_\star \rangle$ for all $\boldsymbol{w} \in \mathbb{R}^d$. This naturally suggests the definitions

$$\mathrm{V}(\boldsymbol{w}) := (1/2)\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2, \quad \mathrm{W}(\boldsymbol{w}) := \langle \nabla F(\boldsymbol{w}) \,|\, \boldsymbol{w} - \boldsymbol{w}_\star \rangle.$$

This yields

$$\|\boldsymbol{h}(\boldsymbol{w})\|^2 = \|\nabla F(\boldsymbol{w})\|^2 \le L_{\nabla F} \mathrm{W}(\boldsymbol{w}) \;,$$

so that **H** 1-*b)* is satisfied with $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1}) = (0, L_{\nabla F})$. **H** 2-*b)* trivially holds with $L_{\mathrm{V}} := 1$ and from

$$\langle \nabla \mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w}) \rangle = -\langle \boldsymbol{w} - \boldsymbol{w}_\star \,|\, \nabla F(\boldsymbol{w}) \rangle = -\mathrm{W}(\boldsymbol{w}),$$

H 2-*c)* is satisfied with $\varrho = 1$. Moreover, we note that the condition (46) reads $\|\boldsymbol{w} - \boldsymbol{w}_\star\| \le c_{\mathrm{V}} \sqrt{\langle \nabla F(\boldsymbol{w}) \,|\, \boldsymbol{w} - \boldsymbol{w}_\star \rangle}$; then $c_{\mathrm{V}} = \infty$ when F is convex but c_{V} is finite when F is strongly convex (see below).

Smooth strongly convex objective functions. We now strengthen the condition on F to strongly convex functions.

CVX2. *The function F is strongly convex with modulus $\mu > 0$.*

In other words, there exists $\mu > 0$ such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ and $\lambda \in (0, 1)$, it holds

$$\lambda F(\boldsymbol{w}) + (1 - \lambda) F(\boldsymbol{w}') \ge F(\lambda \boldsymbol{w} + (1 - \lambda) \boldsymbol{w}') + \lambda(1 - \lambda)(\mu/2) \|\boldsymbol{w} - \boldsymbol{w}'\|^2.$$

Note that a strongly convex function possesses an unique minimizer (see *e.g.,* Corollary 11.17 in [139]), here denoted by $\boldsymbol{w}_\star$. Under **SG**1, **CVX**2 is equivalent to ∇F being μ-strongly monotone *i.e.,* for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$,

$$\langle \nabla F(\boldsymbol{w}') - \nabla F(\boldsymbol{w}) \,|\, \boldsymbol{w}' - \boldsymbol{w} \rangle \ge \mu \|\boldsymbol{w}' - \boldsymbol{w}\|^2 \;, \tag{54}$$

(see *e.g.,* Definition 2.23 and Exercise 17.5 in [139]), which is stronger that (52). As in the convex case above, we may again instantiate the assumptions with

$$\mathrm{V}(\boldsymbol{w}) := (1/2)\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2, \quad \mathrm{W}(\boldsymbol{w}) := \langle \nabla F(\boldsymbol{w}) \,|\, \boldsymbol{w} - \boldsymbol{w}_\star \rangle,$$

Since **CVX**2 implies **CVX** 1, (53) holds and $(c_{h,0}, c_{h,1}) := (0, L_{\nabla F})$. Under **SG** 1, $(\sigma_0^2, \sigma_1^2) := (M^2/n, 0)$. We also have $L_{\mathrm{V}} := 1$ and $\varrho := 1$. Finally, $c_{\mathrm{V}} := \sqrt{2/\mu}$ in (46).

Moreover, an alternative choice for W is possible, *i.e.,*

$$\mathrm{W}(\boldsymbol{w}) = \mathrm{V}(\boldsymbol{w}) := (1/2)\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2.$$

Observe that, from (54),

$$\langle \nabla \mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w}) \rangle = -\langle \boldsymbol{w} - \boldsymbol{w}_\star \,|\, \nabla F(\boldsymbol{w}) \rangle \le -2\mu \mathrm{W}(\boldsymbol{w}) \;.$$

From **SG** 1-*a)* we get $\|\nabla F(\boldsymbol{w})\|^2 \le 2L_{\nabla F}^2 \mathrm{W}(\boldsymbol{w})$. It yields that **H** 1-*b)* is satisfied with $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1}) := (0, 2L_{\nabla F}^2)$. In addition, under **SG** 1, $(\sigma_0^2, \sigma_1^2) := (M^2/n, 0)$. We also have $L_{\mathrm{V}} := 1$ and from (54), **H** 2-*c)* is satisfied with $\varrho := 2\mu$. Finally, $c_{\mathrm{V}} := \sqrt{2}$ in (46).

2) Compressed SA: We show that the assumptions are valid for the methods introduced in Section II-B2. Proofs for this section can be found in Appendix C. We first focus on the compressed SA in (16). Consider the following assumption about the compressor $\boldsymbol{\mathcal{C}}$:

CSA 1. *There exists $\delta_{\boldsymbol{\mathcal{C}}} \in [0, 1]$ such that the compression operator $\boldsymbol{\mathcal{C}}$ satisfies for any $\mathbf{x} \in \mathbb{R}^d$:*

$$\mathbf{E}\left[\|\boldsymbol{\mathcal{C}}(\mathbf{x}, \mathbf{U}) - \mathbf{x}\|^2\right] \le (1 - \delta_{\boldsymbol{\mathcal{C}}}) \|\mathbf{x}\|^2. \tag{55}$$

The Gauss-Southwell estimator (see (12)), obtained by using the projection operator Top_1, satisfies **CSA** 1 with $\delta_{\boldsymbol{\mathcal{C}}} := 1/d$:

Algorithm	$\mathrm{V}(\boldsymbol{w})$	$\mathrm{W}(\boldsymbol{w})$	$\mathrm{V}_\star$	$(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1})$	(τ_0, τ_1)	(σ_0^2, σ_1^2)	L_{V}	ϱ	c_{V}
SG-**SG**1	$F(\boldsymbol{w})$	$\|\nabla F(\boldsymbol{w})\|^2$	$F_\star$	$(0,1)$	$(0,0)$	$(M^2/n, 0)$	$L_{\nabla F}$	1	1
SG-**SG**1-**CVX**1	$\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_\star\|^2$	$\langle \nabla F(\boldsymbol{w}) \,\vert\, \boldsymbol{w}-\boldsymbol{w}_\star\rangle$	0	$(0, L_{\nabla F})$	$(0,0)$	$(M^2/n, 0)$	1	1	∞
SG-**SG**1-**CVX**2	$\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_\star\|^2$	$\langle \nabla F(\boldsymbol{w}) \,\vert\, \boldsymbol{w}-\boldsymbol{w}_\star\rangle$	0	$(0, L_{\nabla F})$	$(0,0)$	$(M^2/n, 0)$	1	1	$1/\sqrt{\mu}$
SG-**SG**1-**CVX**2	$\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_\star\|^2$	$\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_\star\|^2$	0	$(0, 2L_{\nabla F}^2)$	$(0,0)$	$(M^2/n, 0)$	1	2μ	$\sqrt{2}$
Mini-batch EM	$F \circ \mathsf{T}(\boldsymbol{w})$	$\|h(\boldsymbol{w})\|^2$	$F_\star$	$(0,1)$	$(0,0)$	$(\bar{\sigma}_0^2/n\mathrm{m}, \bar{\sigma}_1^2/n\mathrm{m})$	L_{V}	$v_{\min}$	$v_{\max}$
SAEM (i.i.d.)	$F \circ \mathsf{T}(\boldsymbol{w})$	$\|h(\boldsymbol{w})\|^2$	$F_\star$	$(0,1)$	$(0,0)$	$(\bar{\sigma}_0^2/\mathrm{b}_{\mathrm{EM}}, \bar{\sigma}_1^2/\mathrm{b}_{\mathrm{EM}})$	L_{V}	$v_{\min}$	$v_{\max}$
SAEM (IS)	$F \circ \mathsf{T}(\boldsymbol{w})$	$\|h(\boldsymbol{w})\|^2$	$F_\star$	$(0,1)$	(66)	(67)	L_{V}	$v_{\min}$	$v_{\max}$
TD(0)	$\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_\star\|^2$	$\|\boldsymbol{\Phi w}-\boldsymbol{\Phi w}_\star\|^2_{\mathbf{D}_\varpi}$	0	$(0,(1+\lambda)^2)$	$(0,0)$	(70)	1	$1-\lambda$	$1/\sqrt{v_{\min}}$

TABLE III: Required constants for **H** 1 and **H** 2, for each example detailed in Section III-B.

note indeed that $\sum_{i=1}^d x_i^2 \le d(\max_i x_i^2)$ which implies that $\|\mathcal{C}(\mathbf{x},\mathbf{U}) - \mathbf{x}\|^2 = \|\mathbf{x}\|^2 - \max_i x_i^2 \le (1 - 1/d)\|\mathbf{x}\|^2$. The projection operator $\mathrm{Top_h}$ satisfies **CSA** 1 with $(1-\delta_{\mathcal{C}}) := 1 - {}^h/_d$. Other compression operators satisfy this assumption for various constants $\delta_{\mathcal{C}}$, see *e.g.,* [65], [140], [141, Table 1].

We now discuss **H** 1 for the compressed oracles. Since compression can be a random operator, we need to adjust the definition of the filtration $\{\mathcal{F}_k,\ k \in \mathbb{N}\}$ as follows: $\mathcal{F}_0 := \sigma(\boldsymbol{w}_0)$ and for all $k \ge 0$

$$\mathcal{F}_{k+1} := \sigma\left(\boldsymbol{w}_0, \boldsymbol{X}_1, \mathbf{U}_1, \ldots, \boldsymbol{X}_{k+1}, \mathbf{U}_{k+1}\right);$$

$\mathbf{U}_{k+1}$ denotes the random variable $\mathbf{U}$ sampled at iteration $k+1$ when the compression operator $\mathcal{C}$ is applied. With these definitions, $\boldsymbol{w}_k \in \mathcal{F}_k$ for all $k \ge 0$. We assume that the oracle $\boldsymbol{H}$ satisfies **H** 1. We show that the oracle $\mathcal{C}(\boldsymbol{H}(\boldsymbol{w},\boldsymbol{X}),\mathbf{U})$ also satisfies **H** 1 with different constants.

Lemma 1. *Assume that $\boldsymbol{H}$ satisfies **H** 1, with constants $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1})$, $(\tau_{0,k}, \tau_{1,k})$ and (σ_0^2, σ_1^2) and that $\sup_k(\tau_{0,k} + \tau_{1,k}) < \infty$. If $\mathcal{C}$ satisfies **CSA** 1, then, the compressed SG in (16) satisfies **H** 1 with constants in **H** 1 c) and **H** 1 d) given for $\ell \in \{0,1\}$ and $k \ge 0$, by*

$$\begin{aligned}\tau_{\ell,k;\mathcal{C}} &:= ((1+\zeta_1) + (1+\zeta_2)(1+\zeta_1^{-1})(1-\delta_{\mathcal{C}}))\tau_{\ell,k} \\ &\quad + (1+\zeta_2^{-1})(1+\zeta_1^{-1})(1-\delta_{\mathcal{C}})c_{h,\ell} \\ &\quad + (1+\zeta_1^{-1})(1-\delta_{\mathcal{C}})\sigma_\ell^2. \qquad (56)\\ \sigma^2_{\ell;\mathcal{C}} &:= (1-\delta_{\mathcal{C}})\Big((1+\zeta_2)\sup_k \tau_{\ell,k} + (1+\zeta_2^{-1})c_{h,\ell}\Big) \\ &\quad + (1-\delta_{\mathcal{C}})\sigma_\ell^2. \qquad (57)\end{aligned}$$

for any $\zeta_1, \zeta_2 \in \bar{\mathbb{R}}_+$. $c_{h,0}$ and $c_{h,1}$ are unchanged.

An application of Lemma 1 is the Gauss-Southwell update.

Corollary 1. *Consider the Gauss-Southwell update (13), with the corresponding gradient field $\boldsymbol{H}(\boldsymbol{w},\sim) = \mathrm{Top}_1(\nabla \mathrm{F}(\mathrm{w}))$ satisfies $\boldsymbol{H}$ 1 with $(c_{\boldsymbol{h},0;\mathcal{C}}, c_{\boldsymbol{h},1;\mathcal{C}}) = (0,1)$, $(\tau_{0,k;\mathcal{C}}, \tau_{1,k;\mathcal{C}}) = (0, 1 - {}^1/_d)$ and $(\sigma^2_{0;\mathcal{C}}, \sigma^2_{1;\mathcal{C}}) = (0,0)$, for $\mathrm{W} = \|\nabla F(\cdot)\|^2$).*

Indeed, the Top_1 compressor satisfies **CSA** 1 with $(1-\delta_{\mathcal{C}}) = (1 - {}^1/_d)$ and the deterministic field $\boldsymbol{h}(\boldsymbol{w}) \equiv \nabla F(\boldsymbol{w})$ satisfies **H** 1, for $\mathrm{W} = \|\nabla F(\cdot)\|^2$, with $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1}) = (0,1)$, $(\tau_{0,k}, \tau_{1,k}) = (0,0)$ and $(\sigma_0^2, \sigma_1^2) = (0,0)$. Using Lemma 1, with $(\zeta_1, \zeta_2) = (\infty, \infty)$, gives the result.

We note that the compressed random field can be biased even if the original $\boldsymbol{H}$ is not biased: $\tau_{\ell,k} = 0$ does not imply $\tau_{\ell,k;\mathcal{C}} = 0$ if there is compression (*i.e.,* $(1-\delta_{\mathcal{C}}) \neq 0$). As a workaround, a stronger assumption is *unbiased* compression.

CSA 2. *There exists $\omega_{\mathcal{C}} \ge 0$ such that the compression operator $\mathcal{C}$ satisfies for any $\mathbf{x} \in \mathbb{R}^d$:*

$$\mathbb{E}[\mathcal{C}(\mathbf{x},\mathbf{U})] = \mathbf{x}\,, \quad \mathbb{E}\left[\|\mathcal{C}(\mathbf{x},\mathbf{U}) - \mathbf{x}\|^2\right] \le \omega_{\mathcal{C}}\|\mathbf{x}\|^2. \qquad (58)$$

Note that among the operators satisfying this assumption, (scaled) $\mathrm{Rand_h}$ with $1 + \omega_{\mathcal{C}} = {}^d/_h$, (scaled) p-sparsification with $1 + \omega_{\mathcal{C}} = {}^1/_p$, stochastic rounding quantization (15) with $\omega_{\mathcal{C}}$ as a function of Δ [59], [142].

Lemma 2. *Assume that $\boldsymbol{H}$ satisfies **H** 1, with constants $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1})$, $(\tau_{0,k}, \tau_{1,k})$ and (σ_0^2, σ_1^2) and that $\sup_k(\tau_{0,k} + \tau_{1,k}) < \infty$. If $\mathcal{C}$ satisfies **CSA** 2, then, the compressed SA in (16) satisfies **H** 1 with constants in **H** 1 c) and **H** 1 d) given for $\ell \in \{0,1\}$ and $k \ge 0$, by $\tau_{\ell,k;\mathcal{C}} := \tau_{\ell,k}, c_{\ell,k;\mathcal{C}} := c_{\ell,k}$, and*

$$\sigma^2_{\ell;\mathcal{C}} := (1+\omega_{\mathcal{C}})\sigma_\ell^2 + 2\omega_{\mathcal{C}}(c_{\boldsymbol{h},\ell} + \sup_k \tau_{\ell,k}). \qquad (59)$$

Without compression ($\omega_{\mathcal{C}} = 0$), the constants remain unchanged; using compression introduces additional variance. Next, we consider the following result for the compressed SA with perturbed iterate (17). We introduce a third assumption about compression operators that covers the case of a uniformly bounded quantization error in space.

CSA 3. *There exists $\kappa_{\mathcal{C}} \ge 0$ such that the compression operator $\mathcal{C}$ satisfies for any $\mathbf{x} \in \mathbb{R}^d$:*

$$\mathbb{E}\left[\|\mathcal{C}(\mathbf{x},\mathbf{U}) - \mathbf{x}\|^2\right] \le \kappa_{\mathcal{C}}. \qquad (60)$$

Such an assumption is satisfied for operators satisfying **CSA**1 or **CSA**2 on a bounded domain $\mathcal{X} \subset \mathbb{R}^d$, or by using an adaptive number of bits to compress the signal, depending on its scale. This assumption was used, for example, in [143], [144]. For deterministic rounding, as defined in (14), with a quantization step Δ, it holds with $\kappa_{\mathcal{C}} = d\Delta^2$. Such an assumption can be adapted to handle *asynchrony*, for a.s. bounded fields and bounded delays (the compression scheme

can be defined *only* on the points to which it is applied, and depends on the previous sequence of iterates) [76].

Lemma 3. *Assume that $\boldsymbol{H}$ satisfies **H** 1 with constants $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1})$, $(\tau_{0,k}, \tau_{1,k})$ and (σ_0^2, σ_1^2) such that $\sup_k(\tau_{0,k}) < \infty$, for any $k \geq 0$, $\tau_{1,k} = 0$, $\sigma_1^2 = 0$. Assume that $\mathcal{C}$ satisfies **CSA**3, that $\boldsymbol{h}$ is $L_{\boldsymbol{h}}$ Lipschitz, and that $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\cdot, \boldsymbol{X}_{k+1})\right]$ is $L_{\mathbb{E}\boldsymbol{H}}$-Lipschitz. Then the compressed SA in (17) satisfies **H** 1-c) and **H** 1-d), for all $k \geq 0$, for any $\zeta \in \bar{\mathbb{R}}_+$ with $\tau_{1,k;\mathcal{C}} := 0$, $\sigma^2_{1;\mathcal{C}} := 0$, and*

$$\tau_{0,k;\mathcal{C}} := (1+\zeta)\tau_{0,k} + (1+\frac{1}{\zeta})L_{\boldsymbol{h}}^2\kappa_{\mathcal{C}}, \quad \sigma^2_{0;\mathcal{C}} := \sigma_0^2 + L^2_{\mathbb{E}\boldsymbol{H}}\kappa_{\mathcal{C}}.$$

In other words, the SA scheme with perturbed iterates results in an additional bias and variance term. We get the following corollary for compressed SG algorithm STE.

Corollary 2. *Let $\boldsymbol{H}$ be the oracle given by (10) and consider the STE compression SG algorithm (17). Assume **SG**1-a) and $\mathcal{C}$ satisfies **CSA**3. Then the resulting field satisfies **H** 1-c) and **H** 1-d), for all $k \geq 0$, with $\tau_{1,k;\mathcal{C}} := 0$, $\sigma^2_{1;\mathcal{C}} := 0$, and*

$$\tau_{0,k;\mathcal{C}} := L^2_{\nabla F}\kappa_{\mathcal{C}}, \qquad \sigma^2_{0;\mathcal{C}} := \sigma_0^2 + L^2_{\nabla F}\kappa_{\mathcal{C}}.$$

Indeed, we have $\boldsymbol{h} = \nabla F$ and an unbiased stochastic oracle, thus $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\cdot, \boldsymbol{X}_{k+1})\right] = \nabla F$.

Lastly, we study the low-precision SA introduced in (18) where we work with the compressor satisfying:

CSA 4. *There exists $\Delta_{\mathcal{C}} \geq 0$ such that for any $\mathbf{x} \in \mathbb{R}^d$, the compression operator $\mathcal{C}$ satisfies: (i) $\mathbb{E}[\mathcal{C}(\mathbf{x}, \mathbf{U})] = \mathbf{x}$, (ii) denote $\mathcal{B}^d_{\mathcal{C}}$ as the image of $\mathcal{C}(\cdot)$ and it holds for any $\bar{\mathbf{x}} \in \mathcal{B}^d_{\mathcal{C}}$, $\mathcal{C}(\bar{\mathbf{x}}, \mathbf{U}) = \bar{\mathbf{x}}$ almost surely, and for any $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\|\mathcal{C}(\bar{\mathbf{x}} + \mathbf{v}, \mathbf{U}) - \mathbf{x}\|_2^2\right] \leq \|\bar{\mathbf{x}} + \mathbf{v} - \mathbf{x}\|_2^2 + \Delta_{\mathcal{C}}\|\mathbf{v}\|_1. \quad (61)$$

The above assumption is valid for the stochastic rounding quantizer in (15) with $\Delta_{\mathcal{C}} = \Delta$, as shown in [143, Lemma 3] with $\delta = \Delta_{\mathcal{C}}, b = \infty$ therein. Notice that it is also known as a *linear quantizer* in [143].

Lemma 4. *Assume $\boldsymbol{H}$ satisfies **H** 1, with constants $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1})$, $(\tau_{0,k}, \tau_{1,k})$ and (σ_0^2, σ_1^2) and that $\sup_k(\tau_{0,k} + \tau_{1,k}) < \infty$. Consider (18), (19) with constant stepsize $\gamma_k = \bar{\gamma}$ for all k. Assume that $\mathcal{C}$ satisfies **CSA**4. Then, the random field in (19) satisfies **H** 1 with constants in **H** 1 c) and **H** 1 d) given for $\ell \in \{0, 1\}$ and $k \geq 0$, by $\tau_{\ell,k;\mathcal{C}} := \tau_{\ell,k}$ and*

$$\begin{aligned} \sigma^2_{0;\mathcal{C}} &:= \sigma_0^2 + \tfrac{\Delta_{\mathcal{C}}\sqrt{d}}{2\bar{\gamma}}\left(3 + \sup_{k\geq 0}\tau_{0,k} + c_{\boldsymbol{h},0} + \sigma_0^2\right), \\ \sigma^2_{1;\mathcal{C}} &:= \sigma_1^2 + \tfrac{\Delta_{\mathcal{C}}\sqrt{d}}{2\bar{\gamma}}\left(\sup_{k\geq 0}\tau_{1,k} + c_{\boldsymbol{h},1} + \sigma_1^2\right), \end{aligned} \quad (62)$$

while $c_{h,0}$ and $c_{h,1}$ are unchanged.

We note that the random field in (19) inherits the same bias properties from $\boldsymbol{H}$, while its variance is increased to $\mathcal{O}(1 + \frac{\Delta_{\mathcal{C}}\sqrt{d}}{\bar{\gamma}})$. Note that the variance is now inversely proportional to the step size. As we show below in Section IV-B2, such a compressed SA Algorithm converges only to an $\mathcal{O}(\Delta_{\mathcal{C}}\sqrt{d})$ approximate stationary solution.

3) Stochastic EM algorithms: We use the definitions and notations introduced in Section II-B3. The Stochastic EM algorithms solve the root-finding problem for the function

$$\boldsymbol{h}(\boldsymbol{w}) := \bar{\mathsf{s}} \circ \mathsf{T}(\boldsymbol{w}) - \boldsymbol{w}; \quad (63)$$

see (31). Our choice for the Lyapunov function is:

$$\mathrm{V}(\boldsymbol{w}) := F \circ \mathsf{T}(\boldsymbol{w}). \quad (64)$$

In addition to **EM**1 and 2, consider the following assumption

EM 3. *a) There exists $F_\star > -\infty$ such that $F(\boldsymbol{\theta}) \geq F_\star$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$.*

b) The function V is continuously differentiable on $\mathbb{R}^d$ and there exists $L_{\mathrm{V}} \in \mathbb{R}_+$ such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$,

$$\|\nabla\mathrm{V}(\boldsymbol{w}) - \nabla\mathrm{V}(\boldsymbol{w}')\| \leq L_{\mathrm{V}}\|\boldsymbol{w} - \boldsymbol{w}'\|.$$

c) For any $\boldsymbol{w} \in \mathbb{R}^d$, there exists a $d \times d$ positive definite matrix $\mathsf{B}(\boldsymbol{w})$ such that $\nabla\mathrm{V}(\boldsymbol{w}) = -\mathsf{B}(\boldsymbol{w})\boldsymbol{h}(\boldsymbol{w})$. In addition, there exist positive constants $v_{\min} \leq v_{\max}$ such that for any $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$,

$$v_{\min}\|\boldsymbol{w}'\|^2 \leq \|\boldsymbol{w}'\|^2_{\mathsf{B}(\boldsymbol{w})} \leq v_{\max}\|\boldsymbol{w}'\|^2.$$

Lemma 2 in [25] provides sufficient conditions on the regularity of the functions $\boldsymbol{\phi}, \psi, \mathsf{S}_i$ and T, implying that

$$\mathsf{B}(\boldsymbol{w}) = (\nabla\mathsf{T}(\boldsymbol{w}))^\top\, \mathsf{D}_{02}\mathsf{L}(\boldsymbol{w}, \mathsf{T}(\boldsymbol{w}))\, (\nabla\mathsf{T}(\boldsymbol{w})),$$

where $\mathsf{L}(\boldsymbol{w}, \boldsymbol{\theta}) := \psi(\boldsymbol{\theta}) - \langle \boldsymbol{w} \,|\, \boldsymbol{\phi}(\boldsymbol{\theta})\rangle$. Since under **EM** 2, $\mathsf{T}(\boldsymbol{w})$ is the unique minimizer of $\boldsymbol{\theta} \mapsto \mathsf{L}(\boldsymbol{w}, \boldsymbol{\theta})$, $\mathsf{B}(\boldsymbol{w})$ is positive semi-definite. $\nabla\mathsf{T}(\boldsymbol{w})$ is a $d \times d$ matrix and positive definiteness results from conditions on the rank of $\nabla\mathsf{T}(\boldsymbol{w})$.

Lemma 5. *Under **EM** 1 to 3, for any $\boldsymbol{w} \in \mathbb{R}^d$, $\mathrm{V}(\boldsymbol{w}) \geq F_\star$, $\langle \nabla\mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w})\rangle \leq -v_{\min}\|\boldsymbol{h}(\boldsymbol{w})\|^2$, and $\|\nabla\mathrm{V}(\boldsymbol{w})\|^2 \leq v^2_{\max}\|\boldsymbol{h}(\boldsymbol{w})\|^2$. Thus **H** 2 is satisfied with $\mathrm{V}_\star := F_\star$ and*

$$\mathrm{W}(\boldsymbol{w}) := \|\boldsymbol{h}(\boldsymbol{w})\|^2, \quad \varrho := v_{\min}, \quad c_{\mathrm{V}} := v_{\max}; \quad (65)$$

***H** 1-b) is satisfied with $c_{\boldsymbol{h},0} := 0$ and $c_{\boldsymbol{h},1} := 1$.*

Let us now check **H** 1-*c)* and **H** 1 -*d)* for some specific examples of stochastic field $\boldsymbol{H}$.

a) Mini-batch EM: We recall from Section II-B3a the form of mini-batch EM, see eq. (32). We consider the following condition, and obtain the subsequent lemma.

EM 4. *There exist $\bar{\sigma}_0^2, \bar{\sigma}_1^2 \in \mathbb{R}_+$ such that for any $\boldsymbol{w} \in \mathbb{R}^d$*

$$\sup_{i\in\{1,\dots,n\}} \|\bar{\mathsf{s}}_i(\mathsf{T}(\boldsymbol{w}))\|^2 \leq \bar{\sigma}_0^2 + \bar{\sigma}_1^2\mathrm{W}(\boldsymbol{w}).$$

Lemma 6. *Under **EM** 1 to 4, for $\boldsymbol{H}$ given by (32), **H** 1-c) and **H** 1 -d) are verified with $\tau_{\ell,k} = 0$ for all $k \in \mathbb{N}$, and $\sigma_\ell^2 := \bar{\sigma}_\ell^2/\mathsf{b}_{\mathrm{EM}}$, for $\ell \in \{1, 2\}$.*

We refer e.g. to [38, Lemma 7.1.]: first, $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] = \boldsymbol{h}(\boldsymbol{w}_k)$, thus eq. (32) provides an unbiased stochastic oracle. Moreover, the conditional variance of $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ is upper bounded by

$$\tfrac{1}{\mathsf{b}_{\mathrm{EM}}}\tfrac{1}{n}\textstyle\sum_{i=1}^n \|\bar{\mathsf{s}}_i(\mathsf{T}(\boldsymbol{w}_k)) - \frac{1}{n}\sum_{j=1}^n \bar{\mathsf{s}}_j(\mathsf{T}(\boldsymbol{w}_k))\|^2.$$

Second, using $\sum_{i=1}^n \|a_i - n^{-1}\sum_{j=1}^n a_j\|^2 \leq \sum_{i=1}^n \|a_i\|^2$, with **EM**4, **H** 1-*d)* is satisfied with $\sigma_\ell^2 := \bar{\sigma}_\ell^2/\mathsf{b}_{\mathrm{EM}}$.

b) SAEM: We now focus on SAEM, defined in (33). First, consider the case when conditionally to the past, the random variables $\{Z_{i,k+1}^j, 1 \le j \le \mathsf{m}, 1 \le i \le n\}$ are independent, and for all $i \in \{1,\dots,n\}$ and j, the distribution of $Z_{i,k+1}^j$ is $\pi_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$. Then

$$\mathbb{E}^{\mathcal{F}_k}\left[\rho_i^j(\boldsymbol{w}_k)\,\mathsf{S}_i(Z_{i,k+1}^j)\right] = \frac{1}{\mathsf{m}}\bar{\mathsf{s}}_i(\boldsymbol{w}_k)$$

and the SAEM algorithm is an unbiased SA. Moreover, the conditional variance of $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ is equal to

$$\frac{1}{n^2\mathsf{m}}\sum_{i=1}^n \mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{S}_i(Z_{i,k+1}^1) - \frac{1}{n}\sum_{j=1}^n \bar{\mathsf{s}}_j(\boldsymbol{w}_k)\|^2\right].$$

Following the same lines as in the discussions about the Mini-batch EM, we consider the following condition.

EM 5. *There exist constants $\bar\sigma_0^2, \bar\sigma_1^2 \in \mathbb{R}_+$ such that for any $k \in \mathbb{N}$, almost-surely,*

$$\sup_{i\in\{1,\dots,n\}} \mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{S}_i(Z_{i,k+1}^1)\|^2\right] \le \bar\sigma_0^2 + \bar\sigma_1^2 \mathrm{W}(\boldsymbol{w}_k).$$

We have the following lemma.

Lemma 7. *Assume that conditionally to the past, the random variables $\{Z_{i,k+1}^j, 1 \le j \le \mathsf{m}, 1 \le i \le n\}$ are independent, and for all $i \in \{1,\dots,n\}$ and j, $Z_{i,k+1}^j \sim \pi_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$. Assume also **EM** 1 to 3 and 5. Then the oracle given by (33) satisfies **H** 1-c) and **H** 1-d) with $\tau_{\ell,k} = 0$, and $\sigma_\ell^2 := \bar\sigma_\ell^2/(n\mathsf{m})$, for $\ell \in \{1,2\}$ and any $k \in \mathbb{N}$.*

Finally, let us now consider the self-normalized Importance Sampling case; conditionally to the past, the random variables $\{Z_{i,k+1}^j, 1 \le j \le \mathsf{m}, 1 \le i \le n\}$ are independent, and for all $i \in \{1,\dots,n\}$ and j, the distribution of $Z_{i,k+1}^j$ is $\tilde\pi_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$. In that case, SA is not unbiased due to the use of the self-normalized importance weights

$$\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] \neq \boldsymbol{h}(\boldsymbol{w}_k).$$

The expression of the bias and variance is complicated for general functions S_i and for simplicity, we assume here that the functions S_i are bounded (see [127, Theorem 2.3] for an in depth study of self-normalized importance sampling). Define the second moment of the importance ratio with respect to the proposal $\tilde\pi_i(\cdot;\mathsf{T}(\boldsymbol{w}))$

$$\chi_i(\boldsymbol{w}) := \int_{\mathcal{Z}} \left(\frac{\pi_i(z_i;\mathsf{T}(\boldsymbol{w}))}{\tilde\pi_i(z_i;\mathsf{T}(\boldsymbol{w}))}\right)^2 \tilde\pi_i(z_i;\mathsf{T}(\boldsymbol{w}))\mu(\mathrm{d}z_i).$$

Consider the following assumptions.

EM 6. *$s_\star := \max_{1\le i\le n}\sup_{z\in\mathcal{Z}}\|\mathsf{S}_i(z)\|$ is finite and there exist constants $c_{\chi,0}, c_{\chi,1} \in \mathbb{R}_+$ such that for any $\boldsymbol{w} \in \mathbb{R}^d$*

$$\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \chi_i(\boldsymbol{w})\right)^2 \le c_{\chi,0} + c_{\chi,1}\,\mathrm{W}(\boldsymbol{w}).$$

From [127, Theorem 2.1], it holds

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] - \boldsymbol{h}(\boldsymbol{w}_k)\| \le s_\star \frac{12}{\mathsf{m}}\frac{1}{n}\sum_{i=1}^n \chi_i(\boldsymbol{w}),$$

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right]\|^2\right]$$

$$\le s_\star^2 \frac{4}{\mathsf{m}}\frac{1}{n^2}\sum_{i=1}^n \chi_i(\boldsymbol{w}),$$

from which we deduce the following Lemma.

Lemma 8. *Under **EM** 1 to 3 and 6, for $\boldsymbol{H}$ given by (33), $\boldsymbol{H}$ 1-c) and $\boldsymbol{H}$ 1-d) are satisfied with*

$$\tau_{\ell,k} := s_\star^2 \frac{144}{\mathsf{m}^2} c_{\chi,\ell}, \qquad \ell = 0,1 \tag{66}$$

$$\sigma_0^2 := s_\star^2 \frac{4}{n\mathsf{m}}\sqrt{c_{\chi,0} + c_{\chi,1}}, \quad \sigma_1^2 := s_\star^2 \frac{4}{n\mathsf{m}}\sqrt{c_{\chi,1}}. \tag{67}$$

Finally, the case when the samples $\{Z_{i,k+1}^j, 1 \le j \le \mathsf{m}\}$ are the path of an ergodic Markov chain with invariant distribution $\pi_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$, is more complex. We have again

$$\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] \neq \boldsymbol{h}(\boldsymbol{w}_k).$$

An expression of the bias and of the conditional variance of $\boldsymbol{H}$ can be found in [145, Proposition 5] (see also [101, Section 6]) in terms of the iterates of the Markov kernels of the Markov Chain Monte Carlo samplers; these controls rely on Markov Chain theory results (see e.g. [146]) whose exposition is out of the scope of this paper.

4) TD Learning: We use the definitions and notations of Section II-B4. We follow [128] to set up the following assumptions and the missing proofs are relegated to Appendix D.

To design a suitable function W for our assumptions about the SA scheme, we first observe the following lemma from [132, Lemma 4], which shows that the projected Bellman operator $\mathrm{Prj}_\varpi \mathrm{B}$ is a contraction. Note that under **TD** 1, $\mathbf{D}_\varpi$ is positive-definite.

Lemma 9 ([132]). *Assume **TD** 1 and 3. Then, $\mathrm{Prj}_\varpi \mathrm{B}$ is a contraction with respect to $\|\cdot\|_{\mathbf{D}_\varpi}$ with modulus $\lambda \in (0,1)$, i.e., for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, we get*

$$\|\mathrm{Prj}_\varpi \mathrm{B}\,\mathcal{V}_{\boldsymbol{w}} - \mathrm{Prj}_\varpi \mathrm{B}\,\mathcal{V}_{\boldsymbol{w}'}\|_{\mathbf{D}_\varpi} \le \lambda\,\|\mathcal{V}_{\boldsymbol{w}} - \mathcal{V}_{\boldsymbol{w}'}\|_{\mathbf{D}_\varpi}. \tag{68}$$

Therefore, under **TD** 1 and 3, there exists an unique value function $\mathcal{V}_\star$ in $\mathrm{span}(\boldsymbol{\Phi})$ which solves the fixed point of the projected Bellman equation (43): the TD(0) algorithm can be interpreted as a simple SA scheme for solving the projected Bellman fixed point equation. Finally, we set

$$\mathrm{W}(\boldsymbol{w}) := \|\mathcal{V}_{\boldsymbol{w}} - \mathcal{V}_\star\|_{\mathbf{D}_\varpi}^2, \tag{69}$$

which measures the difference between the approximate value function with parameter $\boldsymbol{w}$ and the function $\mathcal{V}_\star$. With the above setup, we are ready to establish **H** 1 under **TD** 1 and 3. It has been checked (see Section II-B4) that under **TD** 3, $\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] = \boldsymbol{h}(\boldsymbol{w}_k)$; hence, **H** 1-c) is satisfied with $\tau_{0,k} = \tau_{1,k} := 0$. Moreover, it can be checked that **H** 1-a) holds, and **H** 1-b) holds with $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1}) = (0, (1+\lambda)^2)$. The other conditions require some technical work, which is summarized below:

Lemma 10. *Assume **TD** 1 to 3. Then, for all $\boldsymbol{w} \in \mathbb{R}^d$ we get $\|\boldsymbol{h}(\boldsymbol{w})\|^2 \le (1+\lambda)^2\,\mathrm{W}(\boldsymbol{w})$ and for any $k \ge 0$, it holds almost-surely,*

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) - \boldsymbol{h}(\boldsymbol{w}_k)\|^2\right] \le \sigma_0^2 + \sigma_1^2 \mathrm{W}(\boldsymbol{w}_k),$$

$$\sigma_0^2 := 6\left(1 + \{\lambda^2+1\}\|\mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}\right), \quad \sigma_1^2 := 2(1+\lambda)^2. \tag{70}$$

We now check **H** 2. Let $\boldsymbol{w}_\star \in \mathbb{R}^d$ be such that $\mathcal{V}_\star = \boldsymbol{\Phi}\boldsymbol{w}_\star$; such a point exists since $\mathcal{V}_\star \in \mathrm{span}(\boldsymbol{\Phi})$. Define

$$\mathrm{V}(\boldsymbol{w}) := (1/2)\|\boldsymbol{w} - \boldsymbol{w}_\star\|^2 . \tag{71}$$

Then **H** 2-*a*), *b*) are verified with $\mathrm{V}_\star := 0$ and $L_\mathrm{V} := 1$. The following Lemma, borrowed from [128, Lemma 3], shows that **H** 2-*c*) is satisfied with $\varrho := 1 - \lambda$.

Lemma 11. *Assume **TD**1 and 3. For any $\boldsymbol{w} \in \mathbb{R}^d$,*

$$\langle \nabla \mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w}) \rangle \le -(1-\lambda)\mathrm{W}(\boldsymbol{w}) ,$$

where $\boldsymbol{h}$ and V are defined in (40) and (71) respectively.

Finally, let us check (46). Since $\mathrm{W}(\boldsymbol{w}) = (\boldsymbol{w} - \boldsymbol{w}_\star)^\top \boldsymbol{\Phi}^\top \mathbf{D}_\varpi \boldsymbol{\Phi}(\boldsymbol{w} - \boldsymbol{w}_\star)$, we get (see Appendix D for a proof)

Lemma 12. *Assume **TD** 1. Then the minimal eigenvalue $\upsilon_{\min}$ of $\boldsymbol{\Phi}^\top \mathbf{D}_\varpi \boldsymbol{\Phi}$ is non-negative and for all $\boldsymbol{w} \in \mathbb{R}^d$, $\sqrt{\upsilon_{\min}}\|\boldsymbol{w}\| \le \|\boldsymbol{w}\|_{\boldsymbol{\Sigma}_\varpi} \le \|\boldsymbol{w}\|$.*

Hence we set $c_\mathrm{V} := 1/\sqrt{\upsilon_{\min}}$ in (46) which can be $+\infty$.

Let us revisit this discussion under stronger conditions on the feature matrix $\boldsymbol{\Phi}$. We assume in the sequel that the number of parameters d needed to approximate the value function $\mathcal{V}_{\boldsymbol{w}} := \boldsymbol{\Phi}\boldsymbol{w}$, is smaller than or equal to the number of states n and that any redundant or irrelevant feature has been removed from the feature matrix $\boldsymbol{\Phi}$ defined in (37). Additionally, we assume that the features are normalized. This is formalized in the following assumption:

TD 4. *The feature matrix is full rank with $\mathrm{rk}(\boldsymbol{\Phi}) = d$. In addition, for all $s \in \mathcal{S}$, $\|\boldsymbol{\phi}(s)\| \le 1$.*

Under **TD**1, 3 and 4, there exists an unique vector $\boldsymbol{w}_\star \in \mathbb{R}^d$ such that $\mathcal{V}_\star = \boldsymbol{\Phi}\boldsymbol{w}_\star$, and $\boldsymbol{w}_\star$ is the unique root to the mean field $\boldsymbol{h}$. Under **TD**4, the *feature covariance matrix*

$$\boldsymbol{\Sigma}_\varpi := \mathbb{E}_\varpi[\boldsymbol{\phi}(S_0)\boldsymbol{\phi}^\top(S_0)] = \boldsymbol{\Phi}^\top \mathbf{D}_\varpi \boldsymbol{\Phi} , \tag{72}$$

is positive-definite: we have an equivalent expression for $\mathrm{W}(\boldsymbol{w})$,

$$\mathrm{W}(\boldsymbol{w}) = \|\boldsymbol{w} - \boldsymbol{w}_\star\|^2_{\boldsymbol{\Sigma}_\varpi} ,$$

and $c_\mathrm{V} = 1/\sqrt{\upsilon_{\min}} > 0$. Finally, (σ_0^2, σ_1^2) are given by (70).

IV. Non-asymptotic convergence bounds

As our first attempt on the theoretical analysis for SA, we overview the non-asymptotic convergence bounds that estimates some properties of the iterates after running the SA scheme for a certain number of iterations. Throughout this section, we focus on the case where the number of iterations of the algorithm is bounded by $T < \infty$, the optimization horizon. Note that the bounds to be presented are in the form of *expected convergence*, where upper bounds on the expected values of the function W are obtained after T iterations.

A. Finite-time bounds and sample complexity of SA

We introduce two simplifying assumptions to streamline our forthcoming discussions.

NA 1. *There exist constants $\tau_0, \tau_1 \in \mathbb{R}_+$ such that for any $k \in \mathbb{N}$, $\tau_{0,k} = \tau_0$, $\tau_{1,k} = \tau_1$.*

NA 2. *It holds $c_\mathrm{V}(\sqrt{\tau_0}/2 + \sqrt{\tau_1}) < \varrho$, where c_V is in (46).*

Recall that c_V can be infinite, and in this case, it is necessary to have $\tau_0 = \tau_1 = 0$ for verifying **NA**2. In words, the above assume that the bias in the oracle $\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})$ is uniformly bounded and small *w.r.t.* the strength of the drift ϱ. Define

$$\mathrm{b}_0 := c_\mathrm{V}\sqrt{\tau_0}/2, \quad \mathrm{b}_1 := c_\mathrm{V}(\sqrt{\tau_0}/2 + \sqrt{\tau_1}). \tag{73}$$

For USO (see Definition 1), $\mathrm{b}_0 = \mathrm{b}_1 = 0$. Set for $\ell \in \{0, 1\}$

$$\begin{aligned}\eta_\ell &:= \sigma_\ell^2 + \tau_\ell + c_{\boldsymbol{h},\ell} + \sqrt{c_{\boldsymbol{h},\ell}}\left(\sqrt{\tau_0} + \sqrt{\tau_1}\right) \\ &\quad + \sqrt{\tau_\ell}\left(\sqrt{c_{\boldsymbol{h},0}} + \sqrt{c_{\boldsymbol{h},1}}\right) ,\end{aligned} \tag{74}$$

$$\gamma_{\max} := 2\{\varrho - \mathrm{b}_1\}/(L_\mathrm{V}\eta_1) , \tag{75}$$

$$\omega_k := 2\{\varrho - \mathrm{b}_1\} - \gamma_k L_\mathrm{V}\eta_1 . \tag{76}$$

If $\eta_1 = 0$, then by convention, $\gamma_{\max} := +\infty$. By construction and under **NA**2, if $\gamma_k \in (0, \gamma_{\max})$, then $\omega_k > 0$.

The essential argument of our analysis is a Robbins-Siegmund type inequality [147], which has played an essential role in the theory of SA since the first works on this subject. Roughly speaking, we control the (expected) changes of the Lyapunov function value $\mathrm{V}(\boldsymbol{w}_{k+1}) - \mathrm{V}(\boldsymbol{w}_k)$ in relation to $\mathrm{W}(\boldsymbol{w}_k)$, the stepsizes, bias, etc.

Lemma 13 (Robbins-Siegmund type inequality)**.** *Assume **H** 1 and 2 and **NA** 1. Then, for any $k \ge 0$, we have almost-surely*

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right] &\le \mathrm{V}(\boldsymbol{w}_k) - (1/2)\gamma_{k+1}\omega_{k+1}\,\mathrm{W}(\boldsymbol{w}_k) \\ &\quad + \gamma_{k+1}\,\mathrm{b}_0 + \gamma_{k+1}^2 L_\mathrm{V}\eta_0/2 .\end{aligned} \tag{77}$$

Proof. Let $k \ge 0$. By **H** 2-*b*), we have

$$\begin{aligned}\mathrm{V}(\boldsymbol{w}_{k+1}) &\le \mathrm{V}(\boldsymbol{w}_k) + \langle \nabla \mathrm{V}(\boldsymbol{w}_k) \,|\, \boldsymbol{w}_{k+1} - \boldsymbol{w}_k \rangle \\ &\quad + (L_\mathrm{V}/2)\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|^2 .\end{aligned} \tag{78}$$

Define $\boldsymbol{b}_k := \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] - \boldsymbol{h}(\boldsymbol{w}_k)$. Computing the conditional expectation of both sides of (78) yields:

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right] &\le \mathrm{V}(\boldsymbol{w}_k) \\ &\quad + \gamma_{k+1}\langle \nabla \mathrm{V}(\boldsymbol{w}_k) \,|\, \boldsymbol{h}(\boldsymbol{w}_k) \rangle + \gamma_{k+1}\langle \nabla \mathrm{V}(\boldsymbol{w}_k) \,|\, \boldsymbol{b}_k \rangle \\ &\quad + \gamma_{k+1}^2 (L_\mathrm{V}/2)\,\mathbb{E}^{\mathcal{F}_k}[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\|^2] .\end{aligned}$$

We now show that

$$|\langle \nabla \mathrm{V}(\boldsymbol{w}_k) \,|\, \boldsymbol{b}_k \rangle| \le \mathrm{b}_0 + \mathrm{b}_1\,\mathrm{W}(\boldsymbol{w}_k), \tag{79}$$

Note first that, using **H** 1-*c*) and (46), we get

$$\begin{aligned}|\langle \nabla \mathrm{V}(\boldsymbol{w}_k) \,|\, \boldsymbol{b}_k \rangle| &\le c_\mathrm{V}\sqrt{\mathrm{W}(\boldsymbol{w}_k)}\{\tau_0 + \tau_1 \mathrm{W}(\boldsymbol{w}_k)\}^{1/2}, \\ &\le c_\mathrm{V}\{\sqrt{\tau_0}\sqrt{\mathrm{W}(\boldsymbol{w}_k)} + \sqrt{\tau_1}\mathrm{W}(\boldsymbol{w}_k)\} ,\end{aligned}$$

and (79) follows from $\sqrt{a} \le (1/2)(1+a)$, $a \ge 0$. Lastly,

$$\begin{aligned}&\mathbb{E}^{\mathcal{F}_k}[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\|^2] = \|\boldsymbol{h}(\boldsymbol{w}_k) + \boldsymbol{b}_k\|^2 \\ &+ \mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_k}[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})]\|^2\right] \\ &\le \eta_0 + \eta_1 \mathrm{W}(\boldsymbol{w}_k) ,\end{aligned}$$

where η_0, η_1 are defined in (74). Combining the results above and **H** 2-*c*), we get (77). □

In Appendix E, we provide a slightly different version for Lemma 13 for the particular case of GD. An important consequence of Lemma 13 is that it allows us to deduce a non-asymptotic bound on $\{\mathrm{W}(\boldsymbol{w}_k)\}_{k=0}^{T-1}$ as follows.

Theorem 1. *Assume **H** 1, 2 and **NA** 1, 2. Assume in addition that the step sizes $\{\gamma_k,\ k \in \mathbb{N}\}$ are chosen such that $\gamma_k \in (0, \gamma_{\max})$. Then, for any $T \geq 1$, we get*

$$\sum_{k=0}^{T-1} \frac{\gamma_{k+1}\omega_{k+1}}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\omega_{\ell+1}}\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] \leq \frac{2(\mathbb{E}[\mathrm{V}(\boldsymbol{w}_0)] - \mathrm{V}_\star) + L_{\mathrm{V}}\eta_0 \sum_{k=0}^{T-1}\gamma_{k+1}^2}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\omega_{\ell+1}} + \frac{2\,\mathrm{b}_0 \sum_{k=0}^{T-1}\gamma_{k+1}}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\omega_{\ell+1}}\,. \tag{80}$$

Proof. Taking the expectations of both sides of (77) gives

$$(1/2)\gamma_{k+1}\omega_{k+1}\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] \leq \mathbb{E}[\mathrm{V}(\boldsymbol{w}_k)] - \mathbb{E}[\mathrm{V}(\boldsymbol{w}_{k+1})] + \gamma_{k+1}\,\mathrm{b}_0 + \gamma_{k+1}^2 L_{\mathrm{V}}\eta_0/2.$$

We obtain (80) by summing these inequalities from $k = 0$ to $k = T-1$ and by using **H** 2-*a*); note that under **NA** 2, $\gamma_{\max} > 0$ and $\gamma_{\ell+1}\omega_{\ell+1} > 0$. □

Suppose that $\mathrm{b}_0 = 0$, under an appropriate stepsize policy (e.g., $\gamma_k = \gamma_{\max}/\sqrt{T}$), it can be shown that the RHS of (80) goes to zero when $T \to \infty$. As discussed in Section III-A, for strict Lyapunov functions, $\mathrm{W}(\boldsymbol{w}) = 0$ implies that $\boldsymbol{w}$ is an equilibrium point to the vector field $\boldsymbol{h}(\boldsymbol{w}) = \boldsymbol{0}$. We further recall that under (50), controlling $\mathbb{E}\,[\mathrm{W}(\boldsymbol{w}_R)]$ at some random stopping time R leads to a high probability bound for the distance from $\boldsymbol{w}_R$ to $\{\boldsymbol{h} = 0\}$.

The following discussions demonstrate how to apply (80) to obtain performance estimates for the SA scheme.

Random Stopping. We first discuss ways to implement the LHS in (80). In particular, we note that it can be viewed as the expected value $\mathbb{E}[\mathrm{W}(\boldsymbol{w}_{R_T})]$, where R_T is a random variable taking values in $\{0, \cdots, T-1\}$, independent of $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$, and with probability mass function

$$\mathbb{P}(R_T = k) = \frac{\gamma_{k+1}\omega_{k+1}}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\omega_{\ell+1}}\,. \tag{81}$$

We may regard $\boldsymbol{w}_{R_T}$ as the output of the SA scheme terminated at the random iteration number R_T. Equivalently, one can look at such a randomization scheme from a slightly different perspective. Here, one can also run the SA algorithm for T iterations, but randomly choose a point $\boldsymbol{w}_{R_T}$ from its trajectory as the output of the algorithm. Clearly, in the latter method, the algorithm only needs to be run for the first R_T iterations. Note, however, that the primary goal of introducing the random iteration number R_T is to derive complexity results, not to save computational effort in the last $T - R_T$ iterations of the algorithm.

Remark 3. *When the function* W *is convex, the LHS of* (80) *is lower bounded by* $\mathbb{E}[\mathrm{W}(\bar{\boldsymbol{w}}_T)]$, *where*

$$\bar{\boldsymbol{w}}_T := \sum_{k=0}^{T-1} \frac{\gamma_{k+1}\,\omega_{k+1}}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\,\omega_{\ell+1}}\boldsymbol{w}_k\,.$$

In this case, one may adopt the convex combination $\bar{\boldsymbol{w}}_T$ to achieve the LHS of (80) *as an alternative to random stopping. Such averaging techniques are referred to as Polyak-Ruppert averaging, and were introduced in [148], [149].*

Constant step size. Let us analyze a special case when a constant stepsize policy is used, *i.e.*, $\gamma_{k+1} = \gamma$ for each $k \in \{0, \ldots, T-1\}$. First, if $\gamma \leq \gamma_{\max}/2$, then (80) can be written more explicitly as follows

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] \leq \mathrm{B} + \frac{2\mathscr{V} + L_{\mathrm{V}}\eta_0 T\gamma^2}{\gamma T\{\varrho - \mathrm{b}_1\}}, \tag{82}$$

where

$$\mathrm{B} := 2\,\mathrm{b}_0/(\varrho - \mathrm{b}_1)\,, \qquad \mathscr{V} := \mathbb{E}[\mathrm{V}(\boldsymbol{w}_0)] - \mathrm{V}_\star. \tag{83}$$

Not surprisingly, the first term, B, in the RHS of (82) which is related to the bias of the random oracle $\boldsymbol{H}$, cannot be made small by any choice of γ. Indeed, according to its definition, b_0 (thus B) scales with τ_0 (see (73)), and τ_0 corresponds (see **H** 1*c*)) to the part of the bias that *does not* scale with the $\boldsymbol{h}$ or W. When $\tau_0 \neq 0$, observing the oracle does not allow to find an exact 0 of the field $\boldsymbol{h}$: indeed, the oracle could for example be equal to $\boldsymbol{h} + \tau_0$ and thus the SA scheme would converge to the roots of $\boldsymbol{h} + \tau_0$, that are distinct from those of $\boldsymbol{h}$.

On the other hand, the second term in the RHS of (82) mixes the dependence on the initial conditions, the bias and variance of the stochastic oracle. It can be adjusted according to the time horizon T and the different parameters of the problem by a suitable choice of γ. The function $\gamma \mapsto 2\mathscr{V}/\gamma + L_{\mathrm{V}}\eta_0 T\gamma$ is minimized on $(0, \gamma_{\max}/2]$ by setting

$$\gamma_T := (2\mathscr{V}/(\eta_0 L_{\mathrm{V}} T))^{1/2} \wedge (\gamma_{\max}/2). \tag{84}$$

Corollary 3. *Assume **H** 1 and 2 and **NA** 1 and 2. Then, for any $T \geq 1$, setting $\gamma_{k+1} := \gamma_T$ for $k \in \{0, \ldots, T-1\}$ we get*

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] \leq \mathrm{B} + \frac{2\sqrt{2\mathscr{V}\eta_0 L_{\mathrm{V}}}}{\sqrt{T}\{\varrho - \mathrm{b}_1\}} \vee \frac{8\mathscr{V}}{\gamma_{\max} T\{\varrho - \mathrm{b}_1\}}\,.$$

In the unbiased case ($\mathrm{B} = 0$), this yields an $\mathcal{O}(1/\sqrt{T})$ convergence rate for SA.

Lower bounds and ϵ-approximate stationarity. In general non-convex optimization, it is intractable to find a global minima of functions or even to test if a point is a local minimum or a high-order saddle point. As a remedy, the most common approach by far is to consider ϵ-approximate stationarity. Our goal is to find a point $\boldsymbol{w}_R \in \mathbb{R}^d$ with

$$\mathbb{E}[\mathrm{W}(\boldsymbol{w}_R)] \leq \epsilon\,, \tag{85}$$

where the expectation is taken over the randomness in both the mean-field oracle and the query R. The use of stationarity as a convergence criterion dates back to the early days of nonlinear optimization (see [150]). Recent years have seen

rapid development of a body of work that studies non-convex optimization through the lens of non-asymptotic convergence rates to ϵ-stationary points [41], [55], [151], [152].

We will discuss how to choose the constant step size γ and the total number of iterations T to guarantee ϵ-approximate stationarity when USO is satisfied. Corollary 3 implies

Corollary 4. *Assume **H** 1, **H** 2-**NA** 1, **NA** 2 and USO (i.e., $\tau_0 = \tau_1 = 0$). Then, for $\epsilon > 0$, the number of iterations to guarantee an ϵ approximate stationary point (85) is lower bounded by*

$$T(\epsilon) = \frac{8\mathscr{V}\eta_0 L_{\mathrm{V}}}{\epsilon^2 \varrho^2} \vee \frac{8\mathscr{V}}{\gamma_{\max}\epsilon\varrho} \,. \tag{86}$$

Proof. This complexity is obtained by choosing the query R to be uniform over $\{0,\dots,T-1\}$, and T such that the upper bound in Corollary 3 is at most ϵ. Under USO, $\mathrm{b}_1 = 0$ and this bound reads $\mathbb{E}[\mathrm{W}(\boldsymbol{w}_R)] \le 2\sqrt{2\mathscr{V}\eta_0 L_{\mathrm{V}}}/(\sqrt{T}\varrho) \vee 8\mathscr{V}/(\gamma_{\max} T\varrho)$. It is easily checked that (86) ensures (85). □

When $T \ge T(\epsilon)$, the algorithm which returns $\boldsymbol{w}_R$ when R is a uniform random variable on $\{0,\dots,T-1\}$, satisfies the ϵ-approximate stationarity condition $\mathbb{E}\left[\mathrm{W}(\boldsymbol{w}_R)\right] \le \epsilon$. The upper bound in Corollary 4 shows that there are two regimes depending on the value of ϵ *w.r.t.* $\gamma_{\max}\eta_0 L_{\mathrm{V}}/\varrho = 2\eta_0/\eta_1$.

a) In the high-precision regime where $\epsilon \in (0, 2\eta_0/\eta_1]$,

$$T(\epsilon) = \frac{8\mathscr{V}L_{\mathrm{V}}}{\varrho^2}\frac{\eta_0}{\epsilon^2}, \tag{87}$$

achieved with a constant step size $\gamma(\epsilon) = \frac{\varrho\epsilon}{2\eta_0 L_{\mathrm{V}}}$.

b) In the low-precision regime where $\epsilon > 2\eta_0/\eta_1$,

$$T(\epsilon) = \frac{4\mathscr{V}L_{\mathrm{V}}}{\varrho^2}\frac{\eta_1}{\epsilon}, \tag{88}$$

achieved with a constant stepsize $\gamma = \gamma_{\max}/2$.

Last iterate or "Random" iterate? The standard analysis holds only for random stopping, or, when W is convex, for a convex combination of the iterates. Most practitioners just use the final iterate of SA instead of randomly selecting a solution $\boldsymbol{w}_{R_T}$ from $\{\boldsymbol{w}_k\}_{k=0}^{T-1}$. In the case of SG for convex functions (the mean field is the gradient of a smooth convex function) and of a stochastic oracle without bias, [153], [154] (see also [155]) show that with a clever choice of piecewise constant step, it is possible to obtain for the last iteration the same decay $1/\sqrt{T}$ as for the randomly stopped estimator (or the averaged one in the case where W is convex, see Remark 3). Unfortunately, this approach is strongly linked to the use of a gradient algorithm and does not extend to SA under the general assumptions we consider.

Another option would be to output a solution $\hat{\boldsymbol{w}}_T$ such that

$$\mathrm{W}(\hat{\boldsymbol{w}}_T) = \min_{k=0,\dots,T-1} \mathrm{W}(\boldsymbol{w}_k) \,. \tag{89}$$

In the case of SG for smooth functions and unbiased oracles, the lower bounds reported for example in [156] show that there is no hope of improving the speed in $1/\sqrt{T}$. Using (89) requires additional computational effort for $\{\mathrm{W}(\boldsymbol{w}_k)\}_{k=0}^{T-1}$. Since in most practical cases $\mathrm{W}(\boldsymbol{w}_k)$ cannot be computed exactly, estimation using Monte Carlo simulations would introduce approximation errors and raise robustness issues. For SG, [12, Section 6.1.1.2] describes a two-stages procedure which runs S times the optimization procedure from the same initial condition. This procedure provides a comparable theoretical guarantee, but is rarely used in practice.

Faster Rate. It is possible to improve Theorem 1 if **H** 1 and 2 hold with $\mathrm{W} = \mathrm{V}$; note that, since $\mathrm{W} \ge 0$ and is null on Λ_{V}, this implies $\mathrm{V}_\star = 0$. We note that this setting applies to the stochastic gradient algorithms with strongly convex objective function and a special case of TD(0) learning. Starting from Lemma 13, taking the expectation of both sides of (77), we get

$$\mathbb{E}[\mathrm{W}(\boldsymbol{w}_{k+1})] \le \lambda_{k+1}\,\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] + b_{k+1} \tag{90}$$

where

$$\lambda_k := 1 - \gamma_k(\varrho - \mathrm{b}_1) + \gamma_k^2 L_{\mathrm{V}}\eta_1/2, \tag{91}$$
$$b_k := \gamma_k\,\mathrm{b}_0 + \gamma_k^2 L_{\mathrm{V}}\eta_0/2\,. \tag{92}$$

A straightforward induction shows that for any $k \in \mathbb{N}$,

$$\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] \le \Lambda_{1:k}\,\mathbb{E}[\mathrm{W}(\boldsymbol{w}_0)] + \textstyle\sum_{j=1}^{k} \Lambda_{j+1:k}\, b_j, \tag{93}$$

where we have set $\Lambda_{k+1:k} := 1$ and for $1 \le j \le k$, $\Lambda_{j:k} := \prod_{i=j}^{k}\lambda_i$. Plugging (92) into (93), we obtain for any $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[\mathrm{W}(\boldsymbol{w}_{k+1})] &\le \Lambda_{1:k+1}\mathbb{E}[\mathrm{W}(\boldsymbol{w}_0)] \\ &\quad + \tfrac{L_{\mathrm{V}}\eta_0}{2}\textstyle\sum_{j=1}^{k+1}\gamma_j^2\Lambda_{j+1:k+1} + \mathrm{b}_0\sum_{j=1}^{k+1}\gamma_j\Lambda_{j+1:k+1}\,.\end{aligned}$$

Lemma 17 provides sharp estimates of $\sum_{j=1}^{k+1}\gamma_j^\ell\Lambda_{j+1:k+1}$, for $\ell \in \{1,2\}$, which leads to the following conclusions.

Theorem 2. *Assume **H** 1, 2 are satisfied with $\mathrm{V} = \mathrm{W}$. Assume in addition **NA** 1, 2, and that the stepsize sequence $\{\gamma_k,\ k \in \mathbb{N}\}$ is non-increasing and chosen such that, for any $k \in \mathbb{N}$, $\gamma_k \le \gamma_{\max}/2$ and*

$$\gamma_k/\gamma_{k+1} \le 1 + \gamma_{k+1}(\varrho - \mathrm{b}_1)/4\,. \tag{94}$$

Then, for any $k \ge 0$, we get

$$\mathbb{E}[\mathrm{W}(\boldsymbol{w}_k)] \le \Lambda_{1:k}\,\mathbb{E}[\mathrm{W}(\boldsymbol{w}_0)] + \frac{2L_{\mathrm{V}}\eta_0}{\varrho - \mathrm{b}_1}\gamma_k + \mathrm{B}\,. \tag{95}$$

Proof. Under (94) and $\gamma_{k+1} \le \gamma_{\max}/2$, we have $\lambda_{k+1} \le 1 - \gamma_{k+1}(\varrho - \mathrm{b}_1)/2$. Using Lemma 17 in the Appendix, with $a := (\varrho - \mathrm{b}_1)/2$ and $b := (\varrho - \mathrm{b}_1)/4$, concludes the proof. □

Condition (94) encompasses constant stepsize policies which are common in the literature. Diminishing stepsize sequences are also a popular choice, e.g., the sequence

$$\gamma_{k+1} := \tilde{\gamma}/(k + 1 + T_0)^\beta \quad \text{for } \beta \in (0,1], \tag{96}$$

satisfies (94) by choosing appropriately $\tilde{\gamma}$ and T_0. The following corollary is obtained by setting $\beta = 1$.

Corollary 5. *Assume **H** 1 and 2 are satisfied with $\mathrm{V} = \mathrm{W}$. Assume in addition **NA** 1 and 2. Let $\tilde{\gamma} \ge 6/(\varrho - \mathrm{b}_1)$ and $T_0 \ge 2\tilde{\gamma}/\gamma_{\max}$; and suppose that $\gamma_{k+1} := \tilde{\gamma}/(k+1+T_0)$, for $k \in \{0,\dots,T-1\}$. Then it holds*

$$\mathbb{E}[\mathrm{W}(\boldsymbol{w}_T)] \le$$

$$\mathrm{B}+\left(\frac{T_0}{T+T_0}\right)^{\frac{\bar{\gamma}(\varrho-\mathrm{b}_1)}{2}}\mathbb{E}\left[\mathrm{W}(\boldsymbol{w}_0)\right]+\frac{2L_{\mathrm{V}}\eta_0\tilde{\gamma}}{(T+T_0)(\varrho-\mathrm{b}_1)}\,. \tag{97}$$

Corollary 5 provides a bound on the *last iterate* of SA. In the upper bound, The second term is $\mathcal{O}(T^{-\alpha})$ for some $\alpha \geq 3$, and the third term is $\mathcal{O}(T^{-1})$. As discussed in [50], [157], for SG this upper bound is tight up to factors independent of T.

B. Application to the examples

1) Stochastic Gradient Method: We apply the results of the previous sections to the case of SG. As there are many results in the literature on this subject, *e.g.,* [12], [50]–[52], [158] and the references therein which offer excellent discussions, the only interest in the following results is to show that we can find "state-of-the-art" results by simply choosing V and W.

Proposition 1. *Assume* ***SG*** *1, the sequence* $\{\gamma_k,\ k\in\mathbb{N}^*\}$ *satisfies* $0<\gamma_k\leq 2/L_{\nabla F}$. *The following holds for* SG (11), *1) For any* $T\in\mathbb{N}$,

$$\sum_{k=0}^{T-1}\frac{(2\gamma_{k+1}-L_{\nabla F}\gamma_{k+1}^2)}{\sum_{\ell=0}^{T-1}(2\gamma_{\ell+1}-L_{\nabla F}\gamma_{\ell+1}^2)}\mathbb{E}[\|\nabla F(\boldsymbol{w}_k)\|^2] \leq \frac{\mathbb{E}[F(\boldsymbol{w}_0)]-F_\star+L_{\nabla F}(M^2/n)\sum_{k=0}^{T-1}\gamma_{k+1}^2}{\sum_{\ell=0}^{T-1}(2\gamma_{\ell+1}-L_{\nabla F}\gamma_{\ell+1}^2)}. \tag{98}$$

2) If ***CVX*** *1 also holds, then for any* $T\in\mathbb{N}$,

$$\mathbb{E}[F(\bar{\boldsymbol{w}}_T)]-F_\star\leq\frac{\mathbb{E}[\|\boldsymbol{w}_0-\boldsymbol{w}_\star\|^2]+(M^2/n)\sum_{k=0}^{T-1}\gamma_{k+1}^2}{\sum_{\ell=0}^{T-1}(2\gamma_{\ell+1}-L_{\nabla F}\gamma_{\ell+1}^2)} \tag{99}$$

where $\bar{\boldsymbol{w}}_T=\sum_{k=0}^{T-1}\frac{(2\gamma_{k+1}-L_{\nabla F}\gamma_{k+1}^2\eta)}{\sum_{\ell=0}^{T-1}(2\gamma_{\ell+1}-L_{\nabla F}\gamma_{\ell+1}^2)}\boldsymbol{w}_k$.

3) If ***CVX*** *2 also holds and* $\gamma_k<\mu/L_{\nabla F}^2$, $\gamma_k/\gamma_{k+1}\leq 1+(\mu/4)\gamma_{k+1}$. *Then, for any* $k\in\mathbb{N}$,

$$\mathbb{E}[\|\boldsymbol{w}_k-\boldsymbol{w}_\star\|^2]\leq\frac{L_{\nabla F}^2(M^2/n)}{\mu}\gamma_k^2 +\prod_{l=1}^{k}\left(1-2\mu\gamma_l+2L_{\nabla F}^2\gamma_l^2\right)\mathbb{E}[\|\boldsymbol{w}_0-\boldsymbol{w}_\star\|^2]\,. \tag{100}$$

Notice that item 1 retrieves the conclusions of [12, Theorem 6.1]; item 2 is classical where precise results are given in [50, Section 2] but these results were known since [159], see [158], [160]; item 3 has a long history, see [50, Section 2].

2) Compressed SA: We describe convergence results obtained for compressed SA. We focus on the case of a stochastic gradient field (10), which has been the most studied. The proofs for this section are given in Appendix F. We first analyze the Gauss-Southwell update (12), which has been studied [22], and is also known as the steepest-coordinate descent [161]. Our results can be naturally extended to the compressed algorithm GD for any compressor that satisfies **CSA** 1.

Proposition 2. *Assume* ***SG****1 and consider a constant sequence* $\{\gamma_k,\ k\in\mathbb{N}^*\}$ *chosen as* $\gamma_k={}^1\!/_{8dL_{\nabla F}}$. *The following holds for the iterates of the Gauss-Southwell algorithm* (12).

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla F(\boldsymbol{w}_k)\|^2]\leq\frac{32d^2L_{\nabla F}(\mathbb{E}[F(\boldsymbol{w}_0)]-F_\star)}{T}\,.$$

This result can be compared with [22, Eq. (7)] for Gauss-Southwell, [162, Theorem 5.3] for GD with relatively bounded errors, both of which give a result under an additional assumption of strong convexity. A similar result, also given in [163, Theorem J.3] for an error-compensated version of the compressed algorithm, shows an increase in rate by a factor $\delta_{\mathcal{C}}^{-1}$. Note that our result has a dependence on d^2 that is suboptimal. A refined version of Lemma 13 is given in Appendix E, which allows us to obtain a bound scaling with d. However, the corresponding Lemma 18 only allows us to improve convergence for (i) gradient methods (stochastic or not), (ii) with bias, and (iii) when we choose $\mathrm{V}=F$. Since our interest is much more general, we omit the refinements of this more detailed analysis.

Second, we consider the case of unbiased compression as in (16) applied to a stochastic gradient descent update. This case has been extensively studied in the communication-constrained distributed optimization community, starting with [59] and with several successors [142]. Combining Lemma 2 and Theorem 1 gives the following result:

Proposition 3. *Assume* ***SG*** *1 and consider unbiased-compressed* SG, *i.e. Equation* (16) *with a* $\mathcal{C}$ *satisfying* ***CSA*** *2 and the sequence* $\{\gamma_k,\ k\in\mathbb{N}^*\}$ *satisfying* $0<\gamma_k\leq 1/(L_{\mathrm{V}}(2\omega_{\mathcal{C}}+1))$. *The following holds for any* $T\in\mathbb{N}$,

$$\sum_{k=0}^{T-1}\frac{(2\gamma_{k+1}-L_{\nabla F}\gamma_{k+1}^2)}{\sum_{\ell=0}^{T-1}(2\gamma_{\ell+1}-L_{\nabla F}\gamma_{\ell+1}^2)}\mathbb{E}[\|\nabla F(\boldsymbol{w}_k)\|^2] \leq \frac{\mathbb{E}[F(\boldsymbol{w}_0)]-F_\star+(M^2/n)L_{\nabla F}(1+\omega_{\mathcal{C}})\sum_{k=0}^{T-1}\gamma_{k+1}^2}{\sum_{\ell=0}^{T-1}(2\gamma_{\ell+1}-L_{\nabla F}\gamma_{\ell+1}^2)}. \tag{101}$$

The unbiased compression leads to a multiplicative increase of the noise level and a reduction of the maximum stepsize by a factor $1+2\omega_{\mathcal{C}}$; see [142, Theorem 9]. Third, we apply Theorem 1 to STE using (3) with deterministic rounding (14).

Proposition 4. *Assume* ***SG*** *1 and consider the STE compression* SG *algorithm* (17). *Assume* $\mathcal{C}$ *is* Q_d *in* (14) *with a quantization step* $\Delta<1/(L_{\nabla F}\sqrt{d})$. *Assume that the sequence* $\{\gamma_k,\ k\in\mathbb{N}^*\}$ *is constant* $\{\gamma_k\}=\gamma/\sqrt{T}$, *with* $0<\gamma\leq 1/L_{\nabla F}$. *The following holds for the iterates:*

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla F(\boldsymbol{w}_k)\|^2]\leq 2L_{\nabla F}\sqrt{d}\Delta +\frac{2(\mathbb{E}[\mathrm{V}(\boldsymbol{w}_0)]-\mathrm{V}_\star)+(M^2/n+3L_{\nabla F}\sqrt{d}\Delta)\gamma^2}{\sqrt{T}\gamma/2}$$

This rate can be compared to [82, Theorem 3] for `BinaryConnect`, although we do not assume a bounded domain: we have a convergence rate of $1/\sqrt{T}$ up to a fixed threshold proportional to $\sqrt{d}\Delta$. Finally, we consider the case of low precision SG in terms of (18). Combining Lemma 4 with (82) (which holds under conditions of Theorem 1) gives:

Proposition 5. *Assume* ***SG*** *1 and consider compressed* SG (18) *with constant stepsize* $\gamma_k=\bar{\gamma}$. *If* $\mathcal{C}$ *satisfies* ***CSA*** *4 and*

$$\bar{\gamma}+\Delta_{\mathcal{C}}\sqrt{d}<2/L_{\nabla F}. \tag{102}$$

Then the following holds for any $T \in \mathbb{N}$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla F(\boldsymbol{w}_k)\|^2] \le \frac{2(\mathbb{E}[F(\boldsymbol{w}_0)]-F_\star)}{\bar\gamma T} + \bar\gamma\frac{M^2 L_{\nabla F}}{n} + \sqrt{d}\Delta_{\mathcal{C}} L_{\nabla F}\left(3+M^2/n\right). \quad (103)$$

Eq. (102) holds provided that $\Delta_{\mathcal{C}} \le 2/L_{\nabla F}\sqrt{d}$. In the case of random quantization, this constraints the resolution of quantizer. Moreover, inserting $\bar\gamma = \mathcal{O}(1/\sqrt{T})$ shows that the first two terms in (103) are $\mathcal{O}(1/\sqrt{T})$, while the last term does not vanish with T. This shows that the compressed SG Algorithm converges to a $\mathcal{O}(\Delta_{\mathcal{C}}\sqrt{d})$ approximate stationary solution, i.e., similar to [143, Theorem 2].

3) Stochastic EM algorithms: The stochastic EM algorithms are used to find a root of the mean-field $\boldsymbol{h}$ defined in (63). We specialize our results to the Mini-batch EM introduced in Section III-B3a and to the SAEM algorithm introduced in Section III-B3b, under the assumptions **EM** 1 to 3.

a) Mini-batch EM: This case is an example of USO (see Definition 1) where $\tau_0 = \tau_1 = 0$. Using Section III-B3, it is easily checked that $\eta_0 = \bar\sigma_0^2/\mathsf{b}_{\mathrm{EM}}$, $\eta_1 = 1+\bar\sigma_1^2/\mathsf{b}_{\mathrm{EM}}$, $\gamma_{\max} = 2v_{\min}/(L_{\mathrm{V}}\eta_1)$, and $\omega_k = 2v_{\min} - \gamma_k L_{\mathrm{V}}\eta_1$. Let $T>0$ be the total number of iterations. Assume that, for all $k \in \{0,\dots,T-1\}$, $\gamma_{k+1} < 2v_{\min}/(L_{\mathrm{V}}\eta_1)$. If the Mini-batch EM is stopped at a random iteration index R_T whose distribution is given in (81), then $\mathbb{E}[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2]$ is upper bounded by Theorem 1 as:

Proposition 6. *Assume **EM**1 to 4. Let $\{\boldsymbol{w}_k,\ k\in\mathbb{N}\}$ be the SA sequence with random oracle (32). Then for any $T\in\mathbb{N}$,*

$$\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \frac{2\mathscr{V} + L_{\mathrm{V}}\bar\sigma_0^2\sum_{k=1}^T\gamma_k^2/\mathsf{b}_{\mathrm{EM}}}{\sum_{k=1}^T\gamma_k(2v_{\min}-\gamma_k L_{\mathrm{V}}(1+\bar\sigma_1^2/\mathsf{b}_{\mathrm{EM}}))},$$

*where R_T is the uniform random variable on $\{0,\dots,T-1\}$ and $\mathscr{V}$, L_{V}, $v_{\min}$ and $\bar\sigma_\ell^2$ are defined by (83), **EM**3 and **EM**4.*

Assume constant step sizes on $\{0,\dots,T-1\}$:

$$\gamma_{k+1} = \gamma := \left(\frac{2\mathscr{V}}{\bar\sigma_0^2 L_{\mathrm{V}}}\frac{\mathsf{b}_{\mathrm{EM}}}{T}\right)^{1/2} \wedge \frac{v_{\min}}{L_{\mathrm{V}}(1+\bar\sigma_1^2/\mathsf{b}_{\mathrm{EM}})},$$

and that R_T is the uniform random variable on $\{0,\dots,T-1\}$. Then, using Corollary 3, we get the following upper bound

$$\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \frac{2\sqrt{2\mathscr{V}\bar\sigma_0^2 L_{\mathrm{V}}}}{\sqrt{\mathsf{b}_{\mathrm{EM}}T}v_{\min}} \vee \frac{8\mathscr{V}L_{\mathrm{V}}(1+\bar\sigma_1^2/\mathsf{b}_{\mathrm{EM}})}{2Tv_{\min}^2}. \quad (104)$$

Let $\epsilon \in \left(0, 2\bar\sigma_0^2/\bar\sigma_1^2\right)$. We discuss how to select the number of iterations T, the step size γ and the size of the mini-batch b_{EM} so that $\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \epsilon$. Assume first that

$$0 < \epsilon \le 2\bar\sigma_0^2/(\mathsf{b}_{\mathrm{EM}}+\bar\sigma_1^2). \quad (105)$$

In this case (see (87)), the number of iterations needed to guarantee an ϵ-approximate stationary point is

$$T(\epsilon,\mathsf{b}_{\mathrm{EM}}) := \frac{8\mathscr{V}L_{\mathrm{V}}}{v_{\min}^2}\frac{\bar\sigma_0^2}{\mathsf{b}_{\mathrm{EM}}\epsilon^2}, \quad (106)$$

which grows as ϵ^{-2} as $\epsilon \downarrow 0^+$. The bound in (106) is achieved by taking a constant step size $\gamma(\epsilon,\mathsf{b}_{\mathrm{EM}}) := v_{\min}\mathsf{b}_{\mathrm{EM}}\epsilon/(2\bar\sigma_0^2 L_{\mathrm{V}})$. We observe that $T(\epsilon,\mathsf{b}_{\mathrm{EM}})$ is inversely proportional to the mini-batch size b_{EM}, while the step size is proportional to b_{EM}. Increasing b_{EM} allows more aggressive step sizes to be used and the number of iterations to be reduced accordingly.

It is interesting to study the impact of the choice of b_{EM} on the computational complexity. Note that the computational cost of mini-batch EM depends on two factors: the evaluation of the functions $\bar{\mathsf{s}}_i$ for the current mini-batch $\boldsymbol{X}_{k+1}$, and the cost of updating the parameter by calling the optimization map T. The cost of evaluating the stochastic oracle is $\mathsf{b}_{\mathrm{EM}}\,\mathrm{cost}_{\bar{\mathsf{s}}} + \mathrm{cost}_{\mathsf{T}}$, and after $T(\epsilon,\mathsf{b}_{\mathrm{EM}})$ iterations, this cost is

$$\mathrm{cost}(\epsilon,\mathsf{b}_{\mathrm{EM}}) := \frac{8\mathscr{V}L_{\mathrm{V}}\bar\sigma_0^2}{v_{\min}^2\epsilon^2}\mathrm{cost}_{\bar{\mathsf{s}}}\left(1+\frac{\mathrm{cost}_{\mathsf{T}}}{\mathrm{cost}_{\bar{\mathsf{s}}}}\frac{1}{\mathsf{b}_{\mathrm{EM}}}\right). \quad (107)$$

If $\mathrm{cost}_{\mathsf{T}}$ is negligible *w.r.t.* $\mathrm{cost}_{\bar{\mathsf{s}}}$, there is no clear incentive to take a mini-batch larger than 1 (to see the interest in using a larger mini-batch, we would have to go much further in evaluating computational cost by considering the possibility of parallelization or multithreading, etc.). However, if $\mathrm{cost}_{\mathsf{T}}$ is taken into account - which is often the case, since the evaluation of T requires solving an optimization program - then it becomes interesting to increase the size of the mini-batch. Since this discussion assumes that (105) holds, the maximum batch size is (up to appropriate roundings):

$$\mathsf{b}_{\mathrm{EM}}(\epsilon) := 2\bar\sigma_0^2/\epsilon - \bar\sigma_1^2.$$

This "optimal" mini-batch size is inversely proportional to the accuracy ϵ, consistent with the fact that using aggressive strategies to reduce the number of iterations is a win. It is interesting to note that the step size $\gamma(\epsilon,\mathsf{b}_{\mathrm{EM}}(\epsilon))$ is equal to $v_{\min}(1-\epsilon\bar\sigma_1^2/(2\bar\sigma_0^2))/L_{\mathrm{V}} > v_{\min}/(2L_{\mathrm{V}})$.

Assume now that b_{EM} is such that $\epsilon \ge 2\bar\sigma_0^2/(\mathsf{b}_{\mathrm{EM}}+\bar\sigma_1^2)$. Using (88), the total number of iterations needed to guarantee an ϵ-approximate stationary point is

$$T(\epsilon,\mathsf{b}_{\mathrm{EM}}) := \frac{4\mathscr{V}L_{\mathrm{V}}}{v_{\min}^2}\frac{1+\bar\sigma_1^2/\mathsf{b}_{\mathrm{EM}}}{\epsilon}, \quad (108)$$

which grows as ϵ^{-1} as $\epsilon \downarrow 0^+$. The bound in (108) is achieved by taking a constant step size $\gamma(\epsilon,\mathsf{b}_{\mathrm{EM}}) := \gamma_{\max}/2 = v_{\min}\mathsf{b}_{\mathrm{EM}}/(L_{\mathrm{V}}(\mathsf{b}_{\mathrm{EM}}+\bar\sigma_1^2))$. After $T(\epsilon,\mathsf{b}_{\mathrm{EM}})$ iterations, the total cost of evaluating the stochastic oracles is

$$\mathrm{cost}(\epsilon,\mathsf{b}_{\mathrm{EM}}) := \frac{4\mathscr{V}L_{\mathrm{V}}}{v_{\min}^2\epsilon}(\mathsf{b}_{\mathrm{EM}}+\bar\sigma_1^2)\,\mathrm{cost}_{\bar{\mathsf{s}}}\left(1+\frac{\mathrm{cost}_{\mathsf{T}}}{\mathrm{cost}_{\bar{\mathsf{s}}}}\frac{1}{\mathsf{b}_{\mathrm{EM}}}\right).$$

This quantity is minimized by $\mathsf{b}_{\mathrm{EM}}(\epsilon) := (2\bar\sigma_0^2/\epsilon - \bar\sigma_1^2)\vee\mathsf{b}_{\mathrm{EM}}^\star$,

$$\mathsf{b}_{\mathrm{EM}}^\star := \sqrt{\bar\sigma_1^2\,\mathrm{cost}_{\mathsf{T}}/\,\mathrm{cost}_{\bar{\mathsf{s}}}}.$$

When $\epsilon \le 2\bar\sigma_0^2/(\mathsf{b}_{\mathrm{EM}}^\star+\bar\sigma_1^2)$, then $\mathsf{b}_{\mathrm{EM}}(\epsilon) = 2\bar\sigma_0^2/\epsilon - \bar\sigma_1^2$ and $\mathrm{cost}(\epsilon,\mathsf{b}_{\mathrm{EM}}(\epsilon))$ is equal to the previous case (see (107)); otherwise, the cost $\mathrm{cost}(\epsilon,\mathsf{b}_{\mathrm{EM}}^\star)$ is lower. See the summary of the cost for minibatch EM in Table IV.

b) SAEM with exact sampling (SAEM-ES): Consider the case when conditionally to the past, the random variables $\{Z_{i,k+1}^j, 1\le j\le \mathsf{m}\}$ are sampled from the distribution $\pi_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$ for all $i\in\{1,\dots,n\}$; and $\{Z_{i,k+1}^j, 1\le j\le \mathsf{m}, i\in\{1,\dots,n\}\}$ are independent. Assume **EM**5.

We have (see Section III-B3b) $\eta_0 = \bar{\sigma}_0^2/(n\mathsf{m})$, $\eta_1 = 1 + \bar{\sigma}_1^2/(n\mathsf{m})$, $\mathrm{b}_0 = \mathrm{b}_1 = 0$, $\gamma_{\max} = 2v_{\min}/(L_{\mathrm{V}}\eta_1)$ and $\omega_k = 2v_{\min} - \gamma_k L_{\mathrm{V}}\eta_1$. From Theorem 1, we obtain the following result.

Proposition 7. *Assume **EM**1 to 3. Let $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ be the SA sequence with random oracle (33). Assume in addition **EM**5. Then for any $T \in \mathbb{N}$,*

$$\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \frac{2\mathscr{V} + L_{\mathrm{V}}\bar{\sigma}_0^2 \sum_{k=1}^T \gamma_k^2/(n\mathsf{m})}{\sum_{k=1}^T \gamma_k(2v_{\min} - \gamma_k L_{\mathrm{V}}(1+\bar{\sigma}_1^2/(n\mathsf{m})))},$$

*where R_T is the uniform random variable on $\{0,\dots,T-1\}$ and $\mathscr{V}$, L_{V}, $v_{\min}$ and $\bar{\sigma}_\ell^2$ are defined by (83), **EM**3 and **EM**5.*

We discuss the case of constant step sizes $\gamma_k = \gamma$. By Corollary 3, we choose

$$\gamma := \left(\frac{2\mathscr{V}}{\bar{\sigma}_0^2 L_{\mathrm{V}}}\frac{n\mathsf{m}}{T}\right)^{1/2} \wedge \frac{v_{\min}}{L_{\mathrm{V}}(1+\bar{\sigma}_1^2/(n\mathsf{m}))},$$

and obtain the upper bound

$$\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \frac{2\sqrt{2\mathscr{V}\bar{\sigma}_0^2 L_{\mathrm{V}}}}{\sqrt{n\mathsf{m}\,T}v_{\min}} \vee \frac{8\mathscr{V}L_{\mathrm{V}}(1+\bar{\sigma}_1^2/(n\mathsf{m}))}{2Tv_{\min}^2},$$

where R_T is the uniform random variable on $\{0,\dots,T-1\}$.

Let $\epsilon \in \left(0, 2\bar{\sigma}_0^2/\bar{\sigma}_1^2\right)$. We discuss how to select the number of iterations T, the step size γ and the number m of Monte Carlo samples so that $\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \epsilon$. Assume that m satisfies

$$0 < \epsilon \le 2\bar{\sigma}_0^2/(n\mathsf{m} + \bar{\sigma}_1^2)\ . \tag{109}$$

In this case (see (87)), the total number of iterations needed to guarantee an ϵ-approximate stationary point is

$$T(\epsilon,\mathsf{m}) := \frac{8\mathscr{V}L_{\mathrm{V}}}{v_{\min}^2}\,\frac{\bar{\sigma}_0^2}{n\mathsf{m}\epsilon^2}\ , \tag{110}$$

which grows as ϵ^{-2} as $\epsilon \downarrow 0^+$. The bound in (110) is achieved by taking a constant step size $\gamma(\epsilon,\mathsf{m}) := v_{\min}n\mathsf{m}\epsilon/(2\bar{\sigma}_0^2 L_{\mathrm{V}})$. The minimal number of iterations $T(\epsilon,\mathsf{m})$ is inversely proportional to the Monte Carlo sample size m, while the step size is proportional to m. Increasing m allows more aggressive step sizes to be used and the number of iterations to be reduced accordingly. As in the Mini-batch EM, let us evaluate the computational cost to understand the impact of this choice on the computational complexity. The computational cost depends on the number of calls to the optimization map T (with cost $\mathrm{cost}_{\mathsf{T}}$) and the cost of approximating each of the n expectations $\bar{\mathsf{s}}_i$ by a Monte Carlo sum with m terms (each term has a cost $\mathrm{cost}_{\mathsf{MC}}$). The cost of evaluating one oracle is $\mathrm{cost}_{\mathsf{T}} + n\mathsf{m}\,\mathrm{cost}_{\mathsf{MC}}$, and after $T(\epsilon,\mathsf{m})$ iterations, this cost is

$$\mathrm{cost}(\epsilon,\mathsf{m}) := \frac{8\mathscr{V}L_{\mathrm{V}}\bar{\sigma}_0^2}{v_{\min}^2\epsilon^2}\,\mathrm{cost}_{\mathsf{T}}\left(\frac{1}{n\mathsf{m}} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}\right)\ . \tag{111}$$

Increasing the number m of Monte Carlo samples reduces the cost due to the optimization step T but will have no impact on the Monte Carlo cost. Since this discussion assumes (109), the maximal Monte Carlo sample size is given (up to appropriate roundings) by $\mathsf{m}(\epsilon) := n^{-1}(2\bar{\sigma}_0^2/\epsilon - \bar{\sigma}_1^2)$: the "optimal" sample size is inversely proportional to the accuracy ϵ. It is interesting to note that the step size $\gamma(\epsilon,\mathsf{m}(\epsilon))$ associated with this choice of sample size is $v_{\min}(1 - \epsilon\bar{\sigma}_1^2/(2\bar{\sigma}_0^2))/L_{\mathrm{V}} > v_{\min}/(2L_{\mathrm{V}})$.

Assume now that m is such that $\epsilon \ge 2\bar{\sigma}_0^2/(n\mathsf{m}+\bar{\sigma}_1^2)$. Using (88), the total number of iterations needed to achieve an ϵ-approximate stationary point is lower bounded by

$$T(\epsilon,\mathsf{m}) := \frac{4\mathscr{V}L_{\mathrm{V}}}{v_{\min}^2}\,\frac{1+\bar{\sigma}_1^2/n\mathsf{m}}{\epsilon}\ , \tag{112}$$

which grows as ϵ^{-1} as $\epsilon \downarrow 0^+$. The bound in (112) is achieved by taking a constant step size $\gamma(\epsilon,\mathsf{m}) := \gamma_{\max}/2 = v_{\min}n\mathsf{m}/(L_{\mathrm{V}}(n\mathsf{m}+\bar{\sigma}_1^2))$. After $T(\epsilon,\mathsf{m})$ iterations, the total cost of evaluating the stochastic oracles is

$$\mathrm{cost}(\epsilon,\mathsf{m}) := \frac{4\mathscr{V}L_{\mathrm{V}}}{v_{\min}^2\epsilon}(n\mathsf{m}+\bar{\sigma}_1^2)\ \mathrm{cost}_{\mathsf{T}}\left(\frac{1}{n\mathsf{m}} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}\right).$$

This quantity is minimal with $\mathsf{m}(\epsilon) := \{n^{-1}(2\bar{\sigma}_0^2/\epsilon - \bar{\sigma}_1^2)\} \vee \mathsf{m}^\star$ where

$$\mathsf{m}^\star := n^{-1}\sqrt{\bar{\sigma}_1^2\,\mathrm{cost}_{\mathsf{T}}\,/\,\mathrm{cost}_{\mathsf{MC}}}.$$

When $\epsilon \le 2\bar{\sigma}_0^2/(n\mathsf{m}^\star + \bar{\sigma}_1^2)$, then $\mathsf{m}(\epsilon) = n^{-1}(2\bar{\sigma}_0^2/\epsilon - \bar{\sigma}_1^2)$ and $\mathrm{cost}(\epsilon,\mathsf{m}(\epsilon))$ is equal to the previous case (see (111)); otherwise, the cost $\mathrm{cost}(\epsilon,\mathsf{m}^\star)$ is lower. See the summary of the cost for SAEM-ES in Table IV.

c) SAEM with self-normalized Importance Sampling (SAEM-IS): Consider the case of SAEM-IS: conditionally to the past, the random variables $\{Z_{i,k+1}^j, 1 \le j \le \mathsf{m}, 1 \le i \le n\}$ are independent, and for all $i \in \{1,\dots,n\}$ and j, the distribution of $Z_{i,k+1}^j$ is $\tilde{\pi}_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$. Assume **EM**6. From Section III-B3b, we have for any Monte Carlo batch size m large enough: $\eta_0 = \bar{\sigma}_0^2/\mathsf{m}$, $\eta_1 = 1 + \bar{\sigma}_1^2/\mathsf{m}$, $\mathrm{b}_0 = c_{\mathrm{b}}/\mathsf{m}$ and b_1 small enough so that $v_{\min} - \mathrm{b}_1 \ge v_{\min}/2$. The exact expressions of $\eta_0, \eta_1, \mathrm{b}_0$ and b_1 in terms of $s_\star$, $c_{\chi,\ell}$, n, m and $v_{\max}$ are given in Appendix G. From Theorem 1, we obtain:

Proposition 8. *Assume **EM** 1 to 3. Let $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ be the SA sequence with random oracle (33). Assume in addition that for all $i \in \{1,\dots,n\}$ and $k \ge 0$, the random variables $\{Z_{i,k+1}^j, j \ge 1\}$ are independent and sampled from the distribution $\pi_i(z_i;\mathsf{T}(\boldsymbol{w}_k))$ and **EM**6 holds. Then for any $T \in \mathbb{N}$,*

$$\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \frac{2c_{\mathrm{b}}\sum_{k=1}^T \gamma_k/\mathsf{m}}{\sum_{\ell=1}^T \gamma_\ell\,(2v_{\min} - \gamma_\ell L_{\mathrm{V}}(1+\bar{\sigma}_1^2/(n\mathsf{m})))} + \frac{2\mathscr{V} + L_{\mathrm{V}}\bar{\sigma}_0^2\sum_{k=1}^T \gamma_k^2/(n\mathsf{m})}{\sum_{k=1}^T \gamma_k(2v_{\min} - \gamma_k L_{\mathrm{V}}(1+\bar{\sigma}_1^2/(n\mathsf{m})))},$$

*where R_T is the uniform random variable on $\{0,\dots,T-1\}$ and $\mathscr{V}$, L_{V} and $v_{\min}$ are defined by (83) and **EM**3 respectively.*

Let us discuss how to choose a constant step size γ, the total number of iterations T, and the Monte Carlo batch size m to satisfy the ϵ-approximate stationary condition. SAEM-IS is not a USO algorithm: the stochastic oracles are biased approximations of the mean field. Consequently, $\mathrm{B} \neq 0$ in Corollary 3 as this term does not depend on the step size γ nor on the number of iterations T. Nevertheless, m goes to infinity, we have $\mathrm{b}_0 \to 0$, $\mathrm{b}_1 \to 0$, which implies that $\mathrm{B} \to 0$. As such, B can be made small by a convenient choice of m.

From Corollary 3, we have

$$\mathbb{E}\left[\|\boldsymbol{h}(\boldsymbol{w}_{R_T})\|^2\right] \le \frac{4c_{\mathrm{b}}}{v_{\min}\mathsf{m}} + \frac{4\sqrt{2\mathscr{V}\bar{\sigma}_0^2 L_{\mathrm{V}}}}{\sqrt{\mathsf{m}}\sqrt{T}v_{\min}} \vee \frac{16\mathscr{V}L_{\mathrm{V}}\eta_1}{Tv_{\min}^2}, \tag{113}$$

where R_T is the uniform random variable on $\{0, \cdots, T-1\}$. Such an upper bound is obtained with a constant step size

$$\gamma := \left(\frac{2\mathscr{V}\mathsf{m}}{\bar{\sigma}_0^2 L_{\mathrm{V}} T}\right)^{1/2} \wedge \frac{v_{\min} - \mathrm{b}_1}{L_{\mathrm{V}}(1 + \bar{\sigma}_1^2/\mathsf{m})}.$$

Choose $\kappa \in (0,1)$ such that for any $\epsilon \in \left(0, 2\bar{\sigma}_0^2/\bar{\sigma}_1^2\right]$,

$$\frac{4c_{\mathrm{b}}}{(1-\kappa)v_{\min}} + \epsilon\bar{\sigma}_1^2 \leq \frac{2\bar{\sigma}_0^2}{\kappa}. \tag{114}$$

Let $\epsilon \in \left(0, 2\bar{\sigma}_0^2/\bar{\sigma}_1^2\right]$. First, we choose m such that $\mathrm{B} = 4c_{\mathrm{b}}/(v_{\min}\mathsf{m}) \leq (1-\kappa)\epsilon$. This yields

$$\mathsf{m} \geq 4c_{\mathrm{b}}/((1-\kappa)v_{\min}\epsilon). \tag{115}$$

Now we choose T and m to make the second term in (113) smaller than $\kappa\epsilon$. As in the comments following Corollary 3, we distinguish two regimes: the second term in (113) is lower than $\kappa\epsilon$ if the number of iterations T is larger than

$$\frac{16\mathscr{V}L_{\mathrm{V}}(1+\bar{\sigma}_1^2/\mathsf{m})}{v_{\min}^2\kappa\epsilon} \vee \frac{32\mathscr{V}\bar{\sigma}_0^2 L_{\mathrm{V}}}{\mathsf{m}v_{\min}^2\kappa^2\epsilon^2},$$

and this bound, seen as a function of ϵ, defines two regimes depending on the value of ϵ *w.r.t.* $2\bar{\sigma}_0^2/(\kappa(\mathsf{m}+\bar{\sigma}_1^2))$.

In the high-precision regime where $\epsilon \in \left(0, 2\bar{\sigma}_0^2/\kappa(\mathsf{m}+\bar{\sigma}_1^2)\right]$, the number of iterations T is lower bounded by

$$T(\epsilon, \mathsf{m}) := \frac{32\, L_{\mathrm{V}}\mathscr{V}\bar{\sigma}_0^2}{v_{\min}^2 \mathsf{m}\kappa^2\epsilon^2}.$$

The step size is $\gamma(\epsilon, \mathsf{m}) := \kappa\epsilon v_{\min}\mathsf{m}/(4L_{\mathrm{V}}\bar{\sigma}_0^2)$. Increasing the number of Monte Carlo points allows more aggressive step sizes and decreases the number of iterations. Nevertheless, it has an impact on the computational cost of the algorithm. The cost, per iteration, is the sum of the optimization cost when computing T (it is denoted by $\mathrm{cost}_{\mathsf{T}}$) and the Monte Carlo cost $n\mathsf{m}\,\mathrm{cost}_{\mathsf{MC}}$ when approximating each of the n expectations $\bar{\mathsf{s}}_i$ with m draws ($\mathrm{cost}_{\mathsf{MC}}$ denotes the cost of one Monte Carlo draw). After $T(\epsilon, \mathsf{m})$ iterations, it is equal to

$$\mathrm{cost}(\epsilon, \mathsf{m}) = \frac{32\, L_{\mathrm{V}}\mathscr{V}\bar{\sigma}_0^2}{v_{\min}^2\kappa^2\epsilon^2}\,\mathrm{cost}_{\mathsf{T}}\left(\frac{1}{\mathsf{m}} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}n\right). \tag{116}$$

There is a gain in increasing m; nevertheless, in this high-precision regime, m is upper bounded by $2\bar{\sigma}_0^2/(\kappa\epsilon) - \bar{\sigma}_1^2$, and it also satisfies (115): the definition of κ (see (114)) allows the choice $\mathsf{m}(\epsilon) = 2\bar{\sigma}_0^2/(\kappa\epsilon) - \bar{\sigma}_1^2$.

In the low-precision regime, which corresponds to $\epsilon \geq 2\bar{\sigma}_0^2/\kappa(\mathsf{m}+\bar{\sigma}_1^2)$, the number of iterations is lower bounded by

$$T(\epsilon, \mathsf{m}) := \frac{16\mathscr{V}L_{\mathrm{V}}(1+\bar{\sigma}_1^2/\mathsf{m})}{v_{\min}^2\kappa\epsilon}$$

and the step size $\gamma(\epsilon, \mathsf{m})$ is larger than $v_{\min}/(2L_{\mathrm{V}}(1+\bar{\sigma}_1^2/\mathsf{m}))$. The computational cost is

$$\mathrm{cost}(\epsilon, \mathsf{m}) := \frac{16\mathscr{V}L_{\mathrm{V}}(1+\bar{\sigma}_1^2/\mathsf{m})}{v_{\min}^2\kappa\epsilon}\,\mathrm{cost}_{\mathsf{T}}\left(1 + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}n\mathsf{m}\right).$$

It is minimized with $\mathsf{m}(\epsilon) = \mathsf{m}^\star \vee (2\bar{\sigma}_0^2/(\kappa\epsilon) - \bar{\sigma}_1^2)$ where

$$\mathsf{m}^\star := \sqrt{\bar{\sigma}_1^2\,\mathrm{cost}_{\mathsf{T}}\,/(n\,\mathrm{cost}_{\mathsf{MC}})}.$$

Stochastic EM Algorithms		Computational Cost
Mini-batch EM	High precision regime $\epsilon \in \left(0, \frac{2\bar{\sigma}_0^2}{\mathrm{b}_{\mathrm{EM}}^\star + \bar{\sigma}_1^2}\right]$	$\frac{8\mathscr{V}L_{\mathrm{V}}\bar{\sigma}_0^2}{v_{\min}^2}\frac{\mathrm{cost}_{\bar{\mathsf{s}}}}{\epsilon}\left(\frac{1}{\epsilon} + \frac{\mathrm{cost}_{\mathsf{T}}}{\mathrm{cost}_{\bar{\mathsf{s}}}}\frac{1}{2\bar{\sigma}_0^2 - \epsilon\sigma_1^2}\right)$
	Low precision regime $\epsilon \in \left[\frac{2\bar{\sigma}_0^2}{\mathrm{b}_{\mathrm{EM}}^\star + \bar{\sigma}_1^2}, \frac{2\bar{\sigma}_0^2}{\bar{\sigma}_1^2}\right)$	$\frac{8\mathscr{V}L_{\mathrm{V}}\bar{\sigma}_0^2}{v_{\min}^2}\frac{\mathrm{b}_{\mathrm{EM}}^\star + \bar{\sigma}_1^2}{2\bar{\sigma}_0^2}\frac{\mathrm{cost}_{\bar{\mathsf{s}}}}{\epsilon}\left(1 + \frac{\mathrm{cost}_{\mathsf{T}}}{\mathrm{cost}_{\bar{\mathsf{s}}}}\frac{1}{\mathrm{b}_{\mathrm{EM}}^\star}\right)$
SAEM-ES	High precision regime $\epsilon \in \left(0, \frac{2\bar{\sigma}_0^2}{n\mathsf{m}^\star + \bar{\sigma}_1^2}\right]$	$\frac{8\mathscr{V}L_{\mathrm{V}}\bar{\sigma}_0^2}{v_{\min}^2\epsilon}\,\mathrm{cost}_{\mathsf{T}}\left(\frac{1}{2\bar{\sigma}_0^2 - \epsilon\bar{\sigma}_1^2} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}\frac{1}{\epsilon}\right)$
	Low precision regime $\epsilon \in \left[\frac{2\bar{\sigma}_0^2}{n\mathsf{m}^\star + \bar{\sigma}_1^2}, \frac{2\bar{\sigma}_0^2}{\bar{\sigma}_1^2}\right)$	$\frac{8\mathscr{V}L_{\mathrm{V}}\bar{\sigma}_0^2}{v_{\min}^2\epsilon}\frac{n\mathsf{m}^\star + \bar{\sigma}_1^2}{2\bar{\sigma}_0^2}\,\mathrm{cost}_{\mathsf{T}}\left(\frac{1}{n\mathsf{m}^\star} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}\right)$
SAEM-IS	High precision regime $\epsilon \in \left(0, \frac{2\bar{\sigma}_0^2}{\kappa(\mathsf{m}^\star + \bar{\sigma}_1^2)}\right]$	$\frac{32\,L_{\mathrm{V}}\mathscr{V}\bar{\sigma}_0^2}{v_{\min}^2\kappa\epsilon}\,\mathrm{cost}_{\mathsf{T}}\left(\frac{1}{2\bar{\sigma}_0^2 - \kappa\epsilon\bar{\sigma}_1^2} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}\frac{n}{\kappa\epsilon}\right)$
	Low precision regime $\epsilon \in \left[\frac{2\bar{\sigma}_0^2}{\kappa(\mathsf{m}^\star + \bar{\sigma}_1^2)}, \frac{2\bar{\sigma}_0^2}{\bar{\sigma}_1^2}\right)$	$\frac{16\mathscr{V}L_{\mathrm{V}}(\mathsf{m}^\star + \bar{\sigma}_1^2)}{v_{\min}^2\kappa\epsilon}\,\mathrm{cost}_{\mathsf{T}}\left(\frac{1}{\mathsf{m}^\star} + \frac{\mathrm{cost}_{\mathsf{MC}}}{\mathrm{cost}_{\mathsf{T}}}n\right)$

TABLE IV: Computation costs to find ϵ-approximate stationary points for stochastic EM. The optimal batch size $\mathrm{b}_{\mathrm{EM}}^\star$ (Mini-batch EM), the Monte Carlo sample size $\mathsf{m}^\star$ (SAEM-ES/SAEM-IS), and the step sizes can be found in Section IV-B3. The low-precision regime does not exist if $\bar{\sigma}_1^2 = 0$.

When $\kappa\epsilon \leq 2\bar{\sigma}_0^2/(\mathsf{m}^\star + \bar{\sigma}_1^2)$, then $\mathsf{m}(\epsilon) = 2\bar{\sigma}_0^2/(\kappa\epsilon) - \bar{\sigma}_1^2$ and the cost $\mathrm{cost}(\epsilon, \mathsf{m}(\epsilon))$ is equal to the previous case (see (116)); otherwise, the cost $\mathrm{cost}(\epsilon, \mathsf{m}^\star)$ is lower. See the summary of the cost for SAEM-IS in Table IV.

4) TD Learning: TD(0) is an example of USO (see Definition 1) with $\tau_0 = \tau_1 = 0$. Using the results of Section III-B4, it is easily checked that $\mathrm{b}_0 = \mathrm{b}_1 = 0$, $\eta_0 = 6(1 + 2\|\mathcal{V}_\star\|_{\mathbf{D}_\varpi}^2)$, $\eta_1 = 3(1+\lambda)^2$, $\gamma_{\max} = 2(1-\lambda)/(3(1+\lambda)^2)$ and $\omega_k = 2(1-\lambda) - 3\gamma_k(1+\lambda)^2$. We obtain the following result from Theorem 1 and Remark 3, which extends [128].

Proposition 9. *Assume **TD** 1 to 3 and* $\sup_s \|\boldsymbol{\phi}(s)\| \leq 1$. *Consider the TD(0) sequence defined in (38). Set* $T > 0$, *and let* $\{\gamma_k,\ k \in \mathbb{N}\}$ *be a sequence such that* $0 < \gamma_k < 2(1-\lambda)/(3(1+\lambda)^2)$. *Set*

$$\bar{\boldsymbol{w}}_T := \sum_{k=1}^{T} \frac{\gamma_k(2(1-\lambda) - 3\gamma_k(1+\lambda)^2)}{\sum_{\ell=1}^{T}\gamma_\ell(2(1-\lambda) - 3\gamma_\ell(1+\lambda)^2)}\boldsymbol{w}_k.$$

Then,

$$\mathbb{E}[\|\mathcal{V}_{\bar{\boldsymbol{w}}_T} - \mathcal{V}_\star\|_{\mathbf{D}_\varpi}^2] \leq \frac{\mathbb{E}[\|\boldsymbol{w}_0 - \boldsymbol{w}_\star\|^2] + 6\{1 + 2\|\mathcal{V}_\star\|_{\mathbf{D}_\varpi}^2\}\sum_{k=1}^{T}\gamma_k^2}{\sum_{\ell=1}^{T}\gamma_\ell(2(1-\lambda) - 3\gamma_\ell(1+\lambda)^2)},$$

where Σ_ϖ *and* σ_0^2 *are given in (72) and (70) respectively.* $\boldsymbol{w}_\star$ *is any solution of* $\mathcal{V}_\star = \boldsymbol{\Phi}\boldsymbol{w}_\star$.

Proposition 9 bounds the mean squared distance between the value function estimator under the averaged iterate $\bar{\boldsymbol{w}}_T$ and the fixed point to the projected Bellman equation (43). The strength of this result is that the step sizes and the bound do not depend on the condition number of the feature covariance

matrix (under **TD**4, $v_{\min} > 0$). Let us first consider the case of constant stepsize policy. Set

$$\gamma_T := \sqrt{\frac{2\mathbb{E}[\|\mathcal{V}_{\boldsymbol{w}_0} - \mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}]}{6(1+2\|\mathcal{V}_\star\|^2_{\mathbf{D}_\varpi})T}} \wedge \frac{1-\lambda}{3(1+\lambda)^2}.$$

Using Corollary 3, we get:

Corollary 6. *Assume **TD**1 to 3. Then, for any $T \geq 1$, setting $\gamma_{k+1} := \gamma_T$ for $k \in \{0, \dots, T-1\}$, we get*

$$\mathbb{E}[\|\mathcal{V}_{\bar{\boldsymbol{w}}_T} - \mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}] \leq \frac{2\sqrt{12\mathscr{V}\{1+2\|\mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}\}}}{\sqrt{T}(1-\lambda)} \vee \frac{12\mathscr{V}(1+\lambda)^2}{T(1-\lambda)^2}$$

where $\mathscr{V} = \mathbb{E}[\|\mathcal{V}_{\boldsymbol{w}_0} - \mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}]$ and $\bar{\boldsymbol{w}}_T := T^{-1}\sum_{k=0}^{T-1} \boldsymbol{w}_k$.

Using Lemma 12, we observe that the Lyapunov functions V and W are equivalent, *i.e.,* for all $\boldsymbol{w} \in \mathbb{R}^d$, we get

$$2\sqrt{v_{\min}}\mathrm{V}(\boldsymbol{w}) \leq \mathrm{W}(\boldsymbol{w}) \leq 2\mathrm{V}(\boldsymbol{w}) \,. \tag{117}$$

The conclusions of Lemmas 10 and 11 may be rewritten as

$$\langle \nabla \mathrm{V}(\boldsymbol{w}) \,|\, \boldsymbol{h}(\boldsymbol{w}) \rangle \leq -2\sqrt{v_{\min}}(1-\lambda)\mathrm{V}(\boldsymbol{w}) \,,$$
$$\|\boldsymbol{h}(\boldsymbol{w})\|^2 \leq 2(1+\lambda)^2 \mathrm{V}(\boldsymbol{w}) \,,$$

showing that **H** 1 and **H** 2 are satisfied with $\mathrm{W} = \mathrm{V}$, $\varrho := 2\sqrt{v_{\min}}(1-\lambda)$, $c_{\boldsymbol{h},0} := 0$, and $c_{\boldsymbol{h},1} := 2(1+\lambda)^2$, σ_0^2 and σ_1^2 given by (70).

We can therefore apply Corollary 5; note that under **TD**1 to 4, $\tau_0 = \tau_1 = 0$ and $c_{\mathrm{V}} < \infty$, **NA**1 and 2 are verified and $\mathrm{b}_0 = \mathrm{b}_1 = 0$. This yields the following result, which gives a convergence rate $O(T^{-1})$. Note that under **TD**4, there exists a unique $\boldsymbol{w}_\star \in \mathbb{R}^d$ such that $\mathcal{V}_\star = \boldsymbol{\Phi}\boldsymbol{w}_\star$.

Corollary 7. *Assume **TD**1 to 4. Let $T \geq 1$ and set $\gamma_{k+1} := \tilde{\gamma}/(k+1+T_0)$ for $k = \{0, \dots, T-1\}$, where $\tilde{\gamma} > 3/\sqrt{v_{\min}}(1-\lambda)$ and $T_0 \geq 2\tilde{\gamma}(1+\lambda)^2/(\sqrt{v_{\min}}(1-\lambda))$. Then*

$$\mathbb{E}\left[\|\boldsymbol{w}_T - \boldsymbol{w}_\star\|^2\right] \leq \left(\frac{T_0}{T+T_0}\right)^{\tilde{\gamma}\sqrt{v_{\min}}(1-\lambda)} \mathbb{E}\left[\|\boldsymbol{w}_0 - \boldsymbol{w}_\star\|^2\right] + \frac{12\tilde{\gamma}}{(T+T_0)\sqrt{v_{\min}}(1-\lambda)}\left(1 + \{\lambda^2+1\}\|\mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}\right).$$

In this case, however, the choice of stepsize $\gamma_k = \tilde{\gamma}/(k+T_0)$, depends on the minimal eigenvalue $v_{\min}$ of the feature covariance matrix through the constants $\tilde{\gamma}$, T_0. These results are in the spirit of the *robust* SA introduced by [50]. With constant stepsize policy, after T iterations, the mean squared distance between the value function estimates under the averaged iterate and the fixed point to the projected Bellman equation decreases at the rate $O(T^{-1/2})$. Of course, this is worse than the rate $O(T^{-1})$ for the decreasing step size SA algorithm. However, the expected error bounds of Corollary 6 are guaranteed regardless of the knowledge of $v_{\min}$. The $\mathcal{O}(T^{-1/2})$ bound still holds with any constant step size $\gamma_k = \tilde{\gamma}/\sqrt{T}$, $k \in \{1, \dots, T\}$ with $\tilde{\gamma} > 0$, an error in the choice of $\tilde{\gamma}$ has a linear effect on the error bound in $\max\{\tilde{\gamma}, \tilde{\gamma}^{-1}\}$. This is to be compared with a potentially catastrophic effect of an appropriate choice of the hyperparameters $\tilde{\gamma}$ and T_0 in Corollary 7. These observation justifies the 'robustness' of the method as "fine tuning" of the step sizes to the objective function is not necessary.

V. Almost-sure convergence

This section overviews the asymptotic convergence analysis of SA scheme where we study the behavior of (4) when the optimization horizon tends to infinity ($k \to \infty$) and we will use decreasing step sizes. At the first glance, these asymptotic convergence results may appear less powerful than the non-asymptotic bounds in Section IV, yet we emphasize that these results are presented in flavor of the *almost-sure* convergence towards one of the equilibrium points set. In contrast, the non-asymptotic bounds only show the convergence towards a *near-equilibrium* point within a finite number of iterations, and the results are often given in expectation. In fact, the first results in SA were obtained in the almost-sure convergence framework in the pioneering works of [1] and [2]; see [13, Chapter 1-2] for a historical introduction. Nevertheless, both types of convergence results are equally important for our understanding of the SA schemes.

A. The ODE method

A powerful method for establishing almost-sure convergence results is the so-called ordinary differential equation (ODE) method, which allows us to relate the almost-sure limit point of SA schemes (see (4)) with the limiting sets of the flow of the autonomous ODE

$$\mathrm{d}\boldsymbol{w}/\mathrm{d}t = \boldsymbol{h}(\boldsymbol{w}). \tag{118}$$

The key element is a detailed analysis of the *flow* associated to vector field $\boldsymbol{h}$. Let $\Phi : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ $(t, \mathbf{y}) \mapsto \Phi(t, \mathbf{y}) = \Phi_t(\mathbf{y})$, be a continuous function. The family $\Phi = (\Phi_t)_{t\in\mathbb{R}}$ is called a flow of $\mathbb{R}^d$, if $\Phi_0 = \mathrm{I}$ and for all $(t, s) \in \mathbb{R}^2$, $\Phi_t \circ \Phi_s = \Phi_{t+s}$. Φ is a semi-flow if we substitute the above $\mathbb{R}$ by $\mathbb{R}_+$. The forward orbit of $\mathbf{y} \in \mathbb{R}^d$ is the set $\mathrm{Orb}^+(\mathbf{y}) = \{\Phi_t(\mathbf{y}) : t \geq 0\}$ and the orbit of $\mathbf{y}$ is $\mathrm{Orb}(\mathbf{y}) = \{\Phi_t(\mathbf{y}) : t \in \mathbb{R}\}$.

The continuous vector field $\boldsymbol{h}$ on $\mathbb{R}^d$ is said to have unique integral curve if there exists a flow $\Phi^{\boldsymbol{h}} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$, $(t, \mathbf{y}) \mapsto \Phi^{\boldsymbol{h}}(t, \mathbf{y}) = \Phi^{\boldsymbol{h}}_t(\mathbf{y})$, which is differentiable with respect to t satisfying, for all $t \in \mathbb{R}$,

$$\mathrm{d}\Phi^{\boldsymbol{h}}_t/\mathrm{d}t = \boldsymbol{h}(\Phi^{\boldsymbol{h}}_t), \quad \Phi^{\boldsymbol{h}}_0(\mathbf{y}) = \mathbf{y} \,.$$

A point $\mathbf{w}_\star \in \mathbb{R}^d$ is an *equilibrium* if $\Phi_t(\mathbf{w}_\star) = \mathbf{w}_\star$ for all $t \in \mathbb{R}$. The set $\mathrm{EQ}(\Phi^{\boldsymbol{h}})$ of the equilibrium point of the flow $\Phi^{\boldsymbol{h}}$ coincides with roots of $\boldsymbol{h}$:

$$\mathrm{EQ}(\Phi^{\boldsymbol{h}}) = \{\mathbf{w}_\star \in \mathbb{R}^d, \boldsymbol{h}(\mathbf{w}_\star) = 0\} \,. \tag{119}$$

A set $\Lambda \subset \mathbb{R}^d$ is *invariant* (resp. *positively invariant*) for the flow Φ if $\Phi_t(\Lambda) \subset \Lambda$ for all $t \in \mathbb{R}$ (resp. for all $t \in \mathbb{R}_+$). In this case, we denote by $\Phi|\Lambda$ the flow (resp. semi-flow) restricted to Λ. If $\mathbf{w}_\star$ is an equilibrium point, then the set $\{\mathbf{w}_\star\}$ is invariant.

The ODE method was introduced by [3] and further refined in [4] (with recursive system identification in mind) and extensively studied thereafter; see, e.g., the books [13], [34], [5] for a comprehensive introduction and further references. The possibility of associating limit points of stochastic approximation procedure and a subset of the family of invariant sets of the flow $\Phi^{\boldsymbol{h}}$ - which includes equilibrium points but might include more "complex " sets - is the main motivation for the

work of [3], which extends older contributions on this subject. Previous results in this field have been established under simple and often unverifiable assumptions about the flow $\Phi^{\boldsymbol{h}}$, for example the existence of a single global "asymptotically stable" equilibrium [1] or of a finite number of equilibrium points whose "basins of attraction" cover the whole space. The success of the ODE method stems from the fact that many results of the rich theory of dynamical systems are readily available; see [164], [13, Chapter 4], and [34, Chapter 2].

We will consider two situations. In the first, the most classical, the bias disappears asymptotically, i.e. $\lim_{k\to\infty} \tau_{\ell,k} = 0$, $\ell = 0, 1$. There is an extensive literature on this subject: we give below a very brief overview of these results, which are mainly inspired by [164], [165]; see also [13], [34]. We then extend these results to the case where the bias does not vanish but is bounded by a sufficiently small constant, by extending the results from [39], [40], [166]. To keep the presentation concise, we report the results with only a sketch of proofs.

B. Limit set of SA with vanishing bias

We consider the following *deterministic* sequence, which is a perturbed Euler discretization of the ODE (118)

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \gamma_{k+1}\{\boldsymbol{h}(\mathbf{w}_k) + \mathbf{u}_{k+1} + \mathbf{b}_{k+1}\}. \tag{120}$$

We first strengthen the conditions on the stepsize sequence:

SA 1. *$\{\gamma_k,\ k \in \mathbb{N}\}$ is a non increasing sequence of positive numbers such that $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\lim_{k\to\infty} \gamma_k = 0$.*

For the mean field $\boldsymbol{h}$, we strengthen **H** 2 used in the non-asymptotic analysis, by assuming that the field $\boldsymbol{h}$ is globally Lipschitz; it guarantees the existence and the uniqueness of the solutions of the ODE (118) which can be extended to $\mathbb{R}$:

SA 2. *The vector field $\boldsymbol{h}$ is Lipschitz continuous,* i.e., *for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, $\|\boldsymbol{h}(\mathbf{w}) - \boldsymbol{h}(\mathbf{w}')\| \le L_{\boldsymbol{h}} \|\mathbf{w} - \mathbf{w}'\|$.*

A key assumption in the analysis (and one of the most annoying one) is that the sequence is bounded.

SA 3. $\sup_{k\in\mathbb{N}} \|\mathbf{w}_k\| < \infty$.

This may seem innocuous since we are interested in convergence, but the stability of SA is a non-trivial issue in general. When **SA** 3 is not satisfied - which unfortunately occurs in many practical examples - a classical approach is to project the iterates on a compact set. The projection then modifies the underlying dynamical system, which becomes a differential inclusion; see [13, Chapters 4-5] for details. Another possibility is to project onto a growing sequence of compact sets; this has been advocated by [167] and recently studied in great detail (with Markovian noise) in [37], [136]. Another solution is to reinitialize the sequence upon crossing a growing sequence of boundaries, as suggested by [168] and further worked out in [135]. Technical details to prove stability under these modifications go beyond this survey.

Define $\mathrm{t}_0 := 0$ and for $k \in \mathbb{N}^*$, $\mathrm{t}_k := \sum_{\ell=1}^{k} \gamma_\ell$. The "inverse" of $k \to \mathrm{t}_k$ is the map $m : \mathbb{R}_+ \to \mathbb{N}$ defined by

$$m(t) := \sup\{k \ge 0 : t \ge \mathrm{t}_k\}.$$

By construction, $m(\mathrm{t}_k) = k$. We denote for $T > 0$, $\mathcal{I}_k(T) := \{k+1, \dots, m(\mathrm{t}_k + T)\}$. We finally assume that the perturbations satisfy the following assumption.

SA 4. *It holds $\lim_{k\to\infty} \mathbf{b}_k = 0$ and that for all $T > 0$,*

$$\lim_{n\to\infty} \sup\left\{\left\|\sum_{\ell=n}^{k-1} \gamma_{\ell+1}\mathbf{u}_{\ell+1}\right\| : k \in \mathcal{I}_n(T)\right\} = 0\,.$$

This condition is often referred as the *asymptotic rate of change* (see *e.g.,* [13, Chapter 5, pp.137-138]). We compare the sequence $\{\mathbf{w}_k,\ k \in \mathbb{N}\}$ with the flow induced by the vector field $\boldsymbol{h}$. For a sequence $\{\mathbf{y}_k,\ k \in \mathbb{N}\}$ in $\mathbb{R}^d$, we define the continuous time affine and piecewise constant interpolated functions $\mathbf{Y}, \bar{\mathbf{Y}} : \mathbb{R}_+ \to \mathbb{R}^d$ respectively by

$$\mathbf{Y}(\mathrm{t}_k + s) = \mathbf{y}_k + s\frac{\mathbf{y}_{k+1} - \mathbf{y}_k}{\mathrm{t}_{k+1} - \mathrm{t}_k}, \text{ and } \bar{\mathbf{Y}}(\mathrm{t}_k + s) = \mathbf{y}_k \tag{121}$$

for all $k \in \mathbb{N}$ and $0 \le s < \gamma_{k+1}$. We define, for $(t, T) \in \mathbb{R}_+^2$,

$$\Delta_{\mathbf{u}}(t, T) := \sup_{0\le s\le T} \left\|\int_t^{t+s} \bar{\mathbf{U}}(s)\,\mathrm{d}s\right\|. \tag{122}$$

Note that **SA** 4 is equivalent to $\lim_{t\to\infty} \Delta_{\mathbf{u}}(t, T) = 0$. With these notations, (120) writes, for $t \ge 0$,

$$\mathbf{W}(t) - \mathbf{W}(0) = \int_0^t \left(\boldsymbol{h}(\bar{\mathbf{W}}(s)) + \bar{\mathbf{U}}(s) + \bar{\mathbf{B}}(s)\right) \mathrm{d}s.$$

The first key result for SA algorithms is the convergence of the linear interpolation to the ODE flow:

Proposition 10 ([164, Proposition 4.1])**.** *Assume **SA** 1 to 4. Then for all $T > 0$,*

$$\lim_{t\to\infty} \sup_{0\le s\le T} \left\|\mathbf{W}(t+s) - \Phi_s^{\boldsymbol{h}}(\mathbf{W}(t))\right\| = 0.$$

Let $\delta > 0$, $T > 0$. A (δ, T)-*pseudo-orbit* for a flow Φ from $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ is a finite sequence of partial trajectories: there exist N and time instants $\{t_i\}_{i=1}^N \subset (0, T]$ and $\mathbf{y}_1, \dots, \mathbf{y}_N \in \mathbb{R}^d$ satisfying $\|\mathbf{y}_0 - \boldsymbol{a}\| < \delta$, $\|\Phi_{t_j}(\mathbf{y}_j) - \mathbf{y}_{j+1}\| \le \delta$ for $j = 0, \dots, k-1$ and $\|\mathbf{y}_k - \boldsymbol{b}\| \le \delta$. We write $\boldsymbol{a} \hookrightarrow_{\delta,T}^{\Phi} \boldsymbol{b}$ if there exists a (δ, T)-pseudo-orbit for the flow Φ from $\boldsymbol{a}$ to $\boldsymbol{b}$. We write $\boldsymbol{a} \hookrightarrow^{\Phi} \boldsymbol{b}$ if $\boldsymbol{a} \hookrightarrow_{\delta,T}^{\Phi} \boldsymbol{b}$ for every $\delta > 0, T > 0$. If $\boldsymbol{a} \hookrightarrow^{\Phi} \boldsymbol{a}$ then $\boldsymbol{a}$ is a *chain recurrent point* for the flow Φ. The set of chain-recurrent points of the flow Φ is denoted $\mathrm{CR}(\Phi)$. It is easy to verify that $\mathrm{CR}(\Phi)$ is a closed positively invariant set and that $\mathrm{EQ}(\Phi) \subset \mathrm{CR}(\Phi)$. When $\mathrm{CR}(\Phi)$ is compact, then it is invariant (see [164, Theorem 5.5]). A subset Λ is said *internally chain-recurrent* if Λ is a nonempty compact invariant set of which every point is chain-recurrent for the restricted flow $\Phi|\Lambda$ (i.e., $\mathrm{CR}(\Phi|\Lambda) = \Lambda$).

We now have all the essential notions to formulate the central result for the convergence of SA sequences. In the form given below, the result is due to [165]; see also [164]. Denote by $\mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$ the limit set of the sequence $\{\mathbf{w}_k,\ k \in \mathbb{N}\}$: $\mathbf{w}_\star \in \mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$, if there exists a sequence $\{n_k,\ k \in \mathbb{N}\}$, satisfying $\lim_{k\to\infty} n_k = +\infty$ and $\lim_{k\to\infty} \mathbf{w}_{n_k} = \mathbf{w}_\star$.

Theorem 3 (after [165, Theorem 1.2])**.** *Let $\{\mathbf{w}_k,\ k \in \mathbb{N}\}$ be the sequence given by (120). Assume **SA** 1 to 4. Then $\mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$ is a connected internally chain-recurrent set for the flow $\Phi^{\boldsymbol{h}}$.*

It is shown in [165, Theorem 1.3] that this result cannot be improved, in the sense that any connected internally chain-recurrent set of the vector field $\boldsymbol{h}$ is the limit set of a sequence (120) which satisfies the assumptions **SA** 1 to 4; see [165, Theorem 1.3]. One must therefore be careful that the set of limit points of sequences (120) are not just the equilibrium of the vector field $\boldsymbol{h}$, but can be a priori larger sets on which the vector field $\boldsymbol{h}$ does not necessarily vanish. Thus, to obtain guarantees to converge only to equilibrium points, one must be able to guarantee that the only connected internally recurrent sets are included in $\{\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{h}(\boldsymbol{w}) = 0\}$, which is, of course, possible but typically requires additional assumptions beyond **SA**2; see [34, Chapter 2, Corollary 4].

The characterization of connected internally chain recurrent sets for the flow $\Phi^{\boldsymbol{h}}$ is simplified when the vector field $\boldsymbol{h}$ is associated to a *Lyapunov function*. A continuous function $\mathrm{V} : \mathbb{R}^d \to \mathbb{R}$ is called a *Lyapunov function* for Λ if the function $t \in \mathbb{R}_+ \to \mathrm{V}(\Phi^{\boldsymbol{h}}_t(\mathbf{y}))$ is constant for $\mathbf{y} \in \Lambda$ and strictly decreasing for $\mathbf{y} \in \mathbb{R}^d \setminus \Lambda$. If Λ coincides with the equilibrium set of the flow associated with the vector field $\boldsymbol{h}$, $\{\mathbf{w} \in \mathbb{R}^d, \boldsymbol{h}(\mathbf{w}) = 0\}$, then V is a *strict Lyapunov function* and the flow $\Phi_{\boldsymbol{h}}$ is said to be *gradient-like*.

Corollary 8 ([164, Proposition 6.4, Corollary 6.6])**.** *Let $\{\mathbf{w}_k,\ k \in \mathbb{N}\}$ be the sequence given by (120). Assume **SA** 1 to 4. Assume in addition that*

a) $\Phi^{\boldsymbol{h}}$ admits a Lyapunov function V for a compact invariant set Λ of the flow $\Phi^{\boldsymbol{h}}$.

b) $\mathrm{V}(\Lambda)$ has an empty interior.

Then $\mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$ is contained in Λ and V is constant on $\mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\})$. If in addition V is a strict Lyapunov function for $\Phi^{\boldsymbol{h}}$, then $\mathrm{Lim}(\{\mathbf{w}_k,\ k \in \mathbb{N}\}) \subset \{\boldsymbol{h} = 0\}$.

We conclude this short presentation with a practical characterization of Lyapunov functions which directly connects almost-sure convergence with the assumption **H** 2 we used to establish the non-asymptotic bounds.

Proposition 11. *Assume that there exists a continuously differentiable function $\mathrm{V} : \mathbb{R}^d \to \mathbb{R}$ such that, for any $\mathbf{w} \in \mathbb{R}^d$, $\langle \nabla \mathrm{V}(\mathbf{w}) \,|\, \boldsymbol{h}(\mathbf{w}) \rangle \leq 0$. Define*

$$\Lambda_{\mathrm{V}} := \{\mathbf{w} \in \mathbb{R}^d, \langle \nabla \mathrm{V}(\mathbf{w}) \,|\, \boldsymbol{h}(\mathbf{w}) \rangle = 0\}. \tag{123}$$

Assume that

a) Λ_{V} is compact and invariant for $\Phi^{\boldsymbol{h}}$.

b) $\mathrm{V}(\Lambda_{\mathrm{V}})$ has an empty interior.

Then, every connected internally chain-recurrent set $L \subset \Lambda_{\mathrm{V}}$ is contained in Λ_{V} and V restricted to L is constant. If in addition $\Lambda_{\mathrm{V}} = \{\boldsymbol{h} = 0\}$ then V is a strict Lyapunov function.

Proof. Note indeed that, by the chain rule, we get,

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{V}(\Phi^{\boldsymbol{h}}_t(\mathbf{y})) = \langle \nabla \mathrm{V}(\Phi^{\boldsymbol{h}}_t(\mathbf{y})) \,|\, \boldsymbol{h}(\Phi^{\boldsymbol{h}}_t(\mathbf{y})) \rangle. \tag{124}$$

If the set Λ_{V} is compact and invariant, then V is a Lyapunov function for Λ_{V}. Indeed, since Λ_{V} is invariant, then for any $\mathbf{y} \in \Lambda_{\mathrm{V}}$, $\Phi^{\boldsymbol{h}}_t(\mathbf{y}) \in \Lambda_{\mathrm{V}}$ for all $t \geq 0$. It follows from (124) that $t \mapsto \mathrm{V}(\Phi^{\boldsymbol{h}}_t(y))$ is constant. If $\mathbf{y} \in \mathbb{R}^d \setminus \Lambda_{\mathrm{V}}$, then for all $t \in \mathbb{R}_+$, $\Phi^{\boldsymbol{h}}_t(\mathbf{y}) \notin \Lambda_{\mathrm{V}}$ (since Λ_{V} is invariant) and (124) shows that $t \mapsto \mathrm{V}(\Phi^{\boldsymbol{h}}_t(\mathbf{y}))$ is strictly decreasing. If the set Λ_{V} (see (123)) is compact and invariant for the flow $\Phi^{\boldsymbol{h}}_t$ associated to the vector-field $\boldsymbol{h}$, and if $\mathrm{V}(\Lambda_{\mathrm{V}})$ has an empty interior, then every connected internally chain recurrent set for the flow is included in Λ_{V}. □

A direct proof of Theorem 3 for the field $\boldsymbol{h}$ satisfying the assumptions of Proposition 11 is in [25] and refined by [135].

Almost-sure convergence. Consider the sequence $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ defined by (4). We first show that under **H** 1-**H** 2 and **SA**1-**SA**2, the assumptions **SA**3-**SA**4 are satisfied with probability 1. To apply the results above, we rewrite (4) as:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \gamma_{k+1}\{\boldsymbol{h}(\boldsymbol{w}_k) + \boldsymbol{u}_{k+1} + \boldsymbol{b}_{k+1}\},$$

where we have set

$$\boldsymbol{u}_{k+1} := \boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right], \tag{125}$$

$$\boldsymbol{b}_{k+1} := \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right] - \boldsymbol{h}(\boldsymbol{w}_k)\,. \tag{126}$$

By construction $\{\boldsymbol{u}_k,\ k \in \mathbb{N}\}$ is a martingale difference sequence adapted to the filtration $\mathcal{F} = \{\mathcal{F}_k,\ k \in \mathbb{N}\}$, where for $k \in \mathbb{N}$, $\mathcal{F}_k := \sigma(\boldsymbol{w}_0, \boldsymbol{X}_\ell, \ell = 1, \ldots, k)$. We can then use Theorem 3 to establish the almost-sure convergence of the sequence $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$. The proof is in Appendix H.

Theorem 4. *Assume that $\sum_{k=0}^{\infty} \gamma_k = \infty$, $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$,*

*i) **H** 1 holds with $\lim_k c_{\mathrm{V}} \tau_{0,k} = \lim_{k\to\infty} c_{\mathrm{V}} \tau_{1,k} = 0$ and $\sum_{k=0}^{\infty} \gamma_k c_{\mathrm{V}} \sqrt{\tau_{0,k}} < \infty$, where c_{V} is in (46).*

*ii) **H** 2 holds with the function V satisfying $\lim_{\|\mathbf{w}\|\to\infty} \mathrm{V}(\mathbf{w}) = \infty$.*

Then, with probability one, the sequence $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ converges, $\lim_{k\to\infty} \mathrm{V}(\boldsymbol{w}_k)$ exists and $\mathrm{Lim}(\{\boldsymbol{w}_k,\ k \in \mathbb{N}\})$ is a connected internally chain recurrent set of the vector-field $\boldsymbol{h}$. If in addition, $\Lambda_{\mathrm{V}} := \{\mathbf{w} \in \mathbb{R}^d, \langle \nabla \mathrm{V}(\mathbf{w}) \,|\, \boldsymbol{h}(\mathbf{w}) = 0 \rangle\}$ is compact and invariant for $\Phi^{\boldsymbol{h}}$ and $\mathrm{V}(\Lambda_{\mathrm{V}})$ has an empty interior, then with probability 1, $\mathrm{Lim}(\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}) \subset \Lambda_{\mathrm{V}}$. Finally, if $\mathbb{E}\left[\mathrm{V}(\boldsymbol{w}_0)\right] < \infty$, then $\sup_k \mathbb{E}\left[\mathrm{V}(\boldsymbol{w}_k)\right] < \infty$.

C. Limit set of SA sequences with bounded bias

We now briefly discuss the behavior of the stochastic approximation when the bias is bounded but does not tend to 0. There are few results in the literature, and all of these results were obtained for stochastic gradient algorithms; see [39], [40], [138], [166] . However, the analysis of the proofs shows that these results also hold for general stochastic approximation algorithms. Presenting these results in details would lead us to introduce a large number of delicate mathematical concepts. Unlike the rest of this tutorial, we present the results in a more informal manner. The proofs are based on the use of differential inclusion methods, which generalize ODEs. The idea is thus to see the bias term as a bounded perturbation of the "unbiased" dynamic; these results build on the work of [169]–[173]. Basically, we replace the recursion (120) by

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \gamma_{k+1}\{\boldsymbol{g}_{k+1} + \mathbf{u}_{k+1}\}, \tag{127}$$

where $\boldsymbol{g}_{k+1} \in \boldsymbol{h}^{\tau_0}_{\mathrm{DI}}(\mathbf{w}_k) := \mathrm{B}(\boldsymbol{h}(\mathbf{w}_k), \tau_0)$, with $\mathrm{B}(\mathbf{y}, \delta) := \{\mathbf{y}', \|\mathbf{y}' - \mathbf{y}\| \leq \delta\}$. To analyse such sequence, we should

replace the ODE by a differential inclusion (DI). $\boldsymbol{h}_{\mathrm{DI}}^{\tau_0} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set valued map in the sense that for each $\mathbf{y} \in \mathbb{R}^d$, we have that $\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}(\mathbf{y})$ is a subset of $\mathbb{R}^d$ (in this case, a ball of radius δ, centered at $\boldsymbol{h}(\mathbf{w})$). To determine the limit points of (127), we introduce the DI

$$\mathrm{d}\mathbf{w}(t)/\mathrm{d}t \in \boldsymbol{h}_{\mathrm{DI}}^{\tau_0}(\mathbf{w}(t)). \tag{128}$$

We say that an absolutely continuous curve (a.c.) $\mathbf{w} : \mathbb{R}_+ \to \mathbb{R}^d$ is a *solution* if (128) holds for almost every $t \in \mathbb{R}_+$.

Various notions of continuity exist for set valued maps. The one that will be important for us is the notion of upper semicontinuity. A set valued map $\mathrm{H} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is upper semi continuous (u.s.c.) at $\mathbf{w} \in \mathbb{R}^d$ if for every U, a neighborhood of $\mathrm{H}(\mathbf{w})$, there is $\delta > 0$ such that

$$\|\mathbf{w}' - \mathbf{w}\| \leq \delta \Longrightarrow \mathrm{H}(\mathbf{w}') \subset U.$$

Under **SA**2, the set-valued map $\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}$ is upper semi continuous (u.s.c.) and it follows from [174, Chapter 2.1] (see also [175, Chapter 4.1]) that

Proposition 12. *Assume **SA**2. For every $\mathbf{y} \in \mathbb{R}^d$, there exists a solution to (128) such that $\mathbf{w}(0) = \mathbf{y}$.*

Denote by $\mathcal{C}^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}$ the set of solutions. The *set-valued semi-flow* $\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}$ associated to the set-valued map $\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}$ is defined by

$$\Phi_t^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}(\mathbf{y}) = \{\mathbf{w}(t), \mathbf{w} \in \mathcal{C}^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}, \mathbf{w}(0) = \mathbf{y}\},\ t \in \mathbb{R}_+ \,. \tag{129}$$

As above, we denote by $\mathbf{W}$ the linear interpolation of the sequence $\mathbf{y}$; see (121). Let $\{s_k,\ k \in \mathbb{N}\}$ be an increasing sequence of positive numbers such that $\lim_{k\to\infty} s_k = \infty$. Denote by $\mathbf{W}_k(t) = \mathbf{W}(s_k + t)$, for $t \in \mathbb{R}^+$. The following result follows from [173, Theorem 3.2] (see also [171, Theorem 4.2]). It shows that the limits points of shifted in time linear interpolation converge to solutions of the differential inclusion (128). The formulation is weaker than the one obtained for the ODE in (10) (the statement differs from [169, Theorem 4.2] which is wrong as it is stated).

Proposition 13. *Assume **SA** 1 to 4. For any increasing sequence $\{n_k,\ k \in \mathbb{N}\}$ there exists a subsequence $\{\tilde{n}_k,\ k \in \mathbb{N}\} \subset \{n_k,\ k \in \mathbb{N}\}$ and an absolutely continuous function $\mathbf{W}_\infty$ such that for any $T > 0$,*

$$\lim_{k\to\infty} \sup_{0\leq s\leq T} \|\mathbf{W}_{\tilde{n}_k}(s) - \mathbf{W}_\infty(s)\| = 0 \,.$$

Moreover, $\mathbf{W}_\infty$ is a solution of the differential inclusion (128).

A set $\Lambda \subset \mathbb{R}^d$ is *invariant* if for every $\mathbf{y} \in \Lambda$ there exists a trajectory $\mathbf{w}$ contained in Λ with $\mathbf{w}(0) = \mathbf{y}$. See [169, Section 3.2] for further discussion. Let $\Lambda \subset \mathbb{R}^d$ be an invariant set of (128) and consider $\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}\|\Lambda$, the set-valued flow $\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}$ restricted to Λ. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, we write $\boldsymbol{a} \hookrightarrow \boldsymbol{b}$ if for every $\delta > 0$ and $T > 0$ there exists an integer $n \in \mathbb{N}$, solutions $\mathbf{y}_1, \ldots, \mathbf{y}_n$ to (128), and real numbers $t_1, t_2, \ldots, t_n$ greater than T such that $\mathbf{y}_i(s) \in \Lambda$ for all $0 \leq s \leq t_i$ and $i = 1, \ldots, n$, $\left\|\mathbf{y}_i(t_i) - \mathbf{y}_{i+1}(0)\right\| \leq \delta$, for all $i = 1, \ldots, n-1$, and $\|\mathbf{y}_1(0) - \boldsymbol{a}\| \leq \delta$ and $\|\mathbf{y}_n(t_n) - \boldsymbol{b}\| \leq \delta$. The sequence $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is called an (δ, T) chain in Λ from $\boldsymbol{a}$ to $\boldsymbol{b}$ for the differential inclusion (128).

A point $\boldsymbol{a}$ is chain recurrent if $\boldsymbol{a} \hookrightarrow \boldsymbol{a}$. The set of chain-recurrent point is denoted $\mathrm{CR}(\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}})$. We now have all the necessary notions to formulate an analogue of Theorem 3.

Theorem 5. *Assume **SA** 1 to 4. Then $\mathrm{Lim}(\mathbf{w})$ is a connected internally chain-recurrent set for the flow $\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}$.*

The last step consists in linking the internally chain-recurrent sets of the $\Phi^{\boldsymbol{h}}$, associated to the ODE (118), to those of the set-valued flow $\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}$, associated to the differential inclusion (128). For this purpose we use a general result, [171, Theorem 3.1] on perturbations of set-valued dynamical systems. This key result is used in the proofs of [40] and [39].

Theorem 6. *Assume **SA** 1 to 4. Let $\mathcal{V}$ be an open neighborhood of $\mathrm{CR}(\Phi^{\boldsymbol{h}})$. Then, there exists $\tau_0^{\max}$ such that, for all $\tau_0 \in [0, \tau_0^{\max})$, $\mathrm{CR}(\Phi^{\boldsymbol{h}_{\mathrm{DI}}^{\tau_0}}) \subset \mathcal{V}$.*

In words, this means that the boundary sets of perturbed recursions are in a neighborhood of the boundary sets of unperturbed recursions, if the perturbation is small enough. The weakness of this result is that it is non-quantitative. It can be made more precise, in the case where the mean field is the gradient of a sufficiently regular function; see [39, Theorem 2.1]. We finally give a simplified version of the previous result which is based on Corollary 8:

Corollary 9. *Assume that V is a strict Lyapunov function for $\Phi^{\boldsymbol{h}}$. Then, for all open neighborhood $\mathcal{V}$ of the equilibrium set $\{\boldsymbol{h} = 0\}$, there exists $\tau_0^{\max}$ such that, for all $\tau_0 \in [0, \tau_0^{\max})$, $\mathrm{Lim}(\mathbf{w}) \subset \mathcal{V}$.*

VI. Variance Reduction

Lastly we review on a recent advance in SA, namely the *variance reduction* technique. These results are motivated by relevant applications in ML and SP, and involve slight modifications to the basic SA scheme (4). Below, we shall introduce the main algorithm and the general theoretical results. The proofs are postponed to Appendix I.

Variance reduction in SA aims to provide a sequence of iterates $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ having smaller variance than a plain SA scheme. We describe a general variance reduction technique for non-gradient SA when the mean field $\boldsymbol{h}$ is a finite sum

$$\boldsymbol{h}(\boldsymbol{w}) := \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{h}_i(\boldsymbol{w}), \quad \boldsymbol{w} \in \mathbb{R}^d. \tag{130}$$

Originally, variance reduction techniques for SA were proposed in the stochastic gradient setting (see the survey paper [176] and the classical references [41], [42], [177]–[180], see also [181]–[185]). These results were later extended to non-gradient SA, in a series of works targeting mostly the stochastic versions of the EM algorithm [26], [32], [33], [38], [43], [104], [186]–[188].

Variance reductions techniques are based on the use of *control variates* (see, for example. [126, chapter 5]). Given an unbiased estimator $\mathbf{U}$ of the unknown quantity $\mathbb{E}[\mathbf{U}]$, a control variate is a centered random variable $\mathbf{V}$, such that the variance of $\mathbf{U}+\mathbf{V}$ is less than the variance of $\mathbf{U}$; such a variate yields an estimate $\mathbf{U} + \mathbf{V}$ of $\mathbb{E}[\mathbf{U}]$ with a smaller variance than the original estimator $\mathbf{U}$. The difficulty is to design such a variable $\mathbf{V}$ with minimal additional computational effort /

Algorithm 1: SA-SPIDER

Data: an initial value $\boldsymbol{w}_{\text{init}}$, the number of inner loops $\mathrm{k_{in}}$ and outer loops $\mathrm{k_{out}}$, a stepsize sequence $\gamma_{t,k+1}$ for $t = 1, \dots, \mathrm{k_{out}}$ and $k = 0, \dots, \mathrm{k_{in}} - 1$

Result: A $\mathbb{R}^d$-valued sequence $\boldsymbol{w}_{t,k+1}$, $t = 1, \dots, \mathrm{k_{out}}$ and $k = 0, \dots, \mathrm{k_{in}} - 1$.

1 $\boldsymbol{w}_{0,\mathrm{k_{in}}} = \boldsymbol{w}_{\text{init}}$;
2 **for** $t = 1, \dots, \mathrm{k_{out}}$ **do**
3 $\boldsymbol{w}_{t,0} = \boldsymbol{w}_{t-1,\mathrm{k_{in}}}$ and $\boldsymbol{w}_{t,-1} = \boldsymbol{w}_{t-1,\mathrm{k_{in}}}$;
4 Sample $\mathcal{B}_{t,0}$, of size b_{vr}, in $\{1, \dots, n\}$ with or without replacement ;
5 Set $\boldsymbol{H}^{\mathrm{vr}}_{t,0} = \boldsymbol{h}(\boldsymbol{w}_{t,0})$;
6 **for** $k = 0, \dots, \mathrm{k_{in}} - 1$ **do**
7 Sample $\mathcal{B}_{t,k+1}$, of size b_{vr}, in $\{1, \dots, n\}$ with or without replacement ;
8 $\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} = \boldsymbol{H}^{\mathrm{vr}}_{t,k} + \mathsf{b}_{\mathrm{vr}}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \{\boldsymbol{h}_i(\boldsymbol{w}_{t,k}) - \boldsymbol{h}_i(\boldsymbol{w}_{t,k-1})\}$;
9 $\boldsymbol{w}_{t,k+1} = \boldsymbol{w}_{t,k} + \gamma_{t,k+1} \boldsymbol{H}^{\mathrm{vr}}_{t,k+1}$;

memory footprint. The mechanism proposed in this section relies on the *Stochastic Path-Integrated Differential EstimatoR* (SPIDER) proposed by [180] and later improved in [41], [42]. It has been shown that the SPIDER method offers optimality guarantees among other variance reduction methods in both the stochastic gradient setting and the stochastic EM setting (see e.g. [42, Table 1] and [156] for the stochastic gradient case and [33, section 6] for stochastic EM).

At iteration $(k+1)$, SPIDER defines an oracle for $\boldsymbol{h}(\boldsymbol{w}_k)$ by using the current estimate $\boldsymbol{H}^{\mathrm{vr}}_k$ of $\boldsymbol{h}(\boldsymbol{w}_{k-1})$ as follows

$$\boldsymbol{H}^{\mathrm{vr}}_{k+1} := \tfrac{1}{\mathsf{b}_{\mathrm{vr}}} \textstyle\sum_{i \in \mathcal{B}_{k+1}} \boldsymbol{h}_i(\boldsymbol{w}_k) + \mathbf{V}_{k+1}$$

where

$$\mathbf{V}_{k+1} := \boldsymbol{H}^{\mathrm{vr}}_k - \tfrac{1}{\mathsf{b}_{\mathrm{vr}}} \textstyle\sum_{i \in \mathcal{B}_{k+1}} \boldsymbol{h}_i(\boldsymbol{w}_{k-1}), \tag{131}$$

and $\mathcal{B}_{k+1}$ is a set of indices of cardinality b_{vr} chosen randomly in $\{1, \dots, n\}$ with or without replacement and independently of the history of the algorithm $\mathcal{F}_k$. $\mathbf{V}_{k+1}$ is a random variable chosen to be correlated with the naive oracle $\mathsf{b}_{\mathrm{vr}}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \boldsymbol{h}_i(\boldsymbol{w}_k)$, the correlation relying essentially on the random minibatch $\mathcal{B}_{k+1}$. Since $\mathbb{E}^{\mathcal{F}_k}\big[\mathsf{b}_{\mathrm{vr}}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \boldsymbol{h}_i(\boldsymbol{w}_{k-1})\big] = \boldsymbol{h}(\boldsymbol{w}_{k-1})$, (131) shows that $\mathbf{V}_{k+1}$ is the difference of two estimators of $\boldsymbol{h}(\boldsymbol{w}_{k-1})$; yet its (conditional) expected value is not zero. Namely, we have

$$\mathbb{E}^{\mathcal{F}_k}\big[\mathbf{V}_{k+1}\big] = \boldsymbol{H}^{\mathrm{vr}}_k - \boldsymbol{h}(\boldsymbol{w}_{k-1});$$

(see Lemma 14). The *bias* is tamed by resetting the control variate: every $\mathrm{k_{in}}$ iterations, the control variate is set to zero.

The *SA-SPIDER* algorithm is given by Algorithm 1. The iteration index is a pair (t, k), where t counts the number of control variate updates (*outer loops* number) and k is the number of SA updates since last reset (*inner loops* number).

The following intermediate results show how the bias of the stochastic mean field $\boldsymbol{H}^{\mathrm{vr}}_{\cdot}$, its conditional variance and the L^2-moment of the error $\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})$ evolve along the inner loops. Define the filtration

$$\mathcal{F}_{t,k} := \sigma\left(\boldsymbol{w}_{\text{init}}, \mathcal{B}_{\tau,\kappa}, 1 \le \tau \le t, 1 \le \kappa \le k\right),$$

for all $t \in \{1, \dots, \mathrm{k_{out}}\}$ and $k \in \{1, \dots, \mathrm{k_{in}}\}$. Assume

VR 1. *For all $i \in \{1, \dots, n\}$, there exists $L_i \ge 0$ such that for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$, $\|\boldsymbol{h}_i(\boldsymbol{w}) - \boldsymbol{h}_i(\boldsymbol{w}')\| \le L_i \|\boldsymbol{w} - \boldsymbol{w}'\|$.*

Lemma 14. *Consider the iterates from Algorithm 1. For any $t \in \{1, \dots, \mathrm{k_{out}}\}$ and $k \in \{0, \dots, \mathrm{k_{in}} - 1\}$, it holds*

$$\mathbb{E}^{\mathcal{F}_{t,k}}\big[\boldsymbol{H}^{\mathrm{vr}}_{t,k+1}\big] - \boldsymbol{h}(\boldsymbol{w}_{t,k}) = \boldsymbol{H}^{\mathrm{vr}}_{t,k} - \boldsymbol{h}(\boldsymbol{w}_{t,k-1}) \tag{132}$$

$$\mathbb{E}^{\mathcal{F}_{t,0}}\big[\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\big] = \boldsymbol{H}^{\mathrm{vr}}_{t,0} - \boldsymbol{h}(\boldsymbol{w}_{t,-1}) = 0. \tag{133}$$

Equation (132) shows that at each inner iteration, the oracle $\boldsymbol{H}^{\mathrm{vr}}_{t,k+1}$ is a biased approximation of the mean field $\boldsymbol{h}(\boldsymbol{w}_{t,k})$ and the bias propagates along inner iterations. Conditionally to the initialization of the inner iteration, the bias is equal to the error $\boldsymbol{H}^{\mathrm{vr}}_{t,0} - \boldsymbol{h}(\boldsymbol{w}_{t,-1})$ (see (133)). The strategy used in Line 3 and Line 5 of Algorithm 1 implies that $\boldsymbol{H}^{\mathrm{vr}}_{t,0} - \boldsymbol{h}(\boldsymbol{w}_{t,-1}) = 0$. In that sense, we say that at the beginning of each inner loop, the bias is canceled and the control variate is reset.

Lemma 15. *Assume **VR** 1 and set $L^2 := n^{-1} \sum_{i=1}^n L_i^2$. For any $t \in \{1, \dots, \mathrm{k_{out}}\}$ and $k \in \{0, \dots, \mathrm{k_{in}} - 1\}$, it holds*

$$\mathbb{E}^{\mathcal{F}_{t,k}}\big[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \mathbb{E}^{\mathcal{F}_{t,k}}\big[\boldsymbol{H}^{\mathrm{vr}}_{t,k+1}\big]\|^2\big] \le \frac{L^2}{\mathsf{b}_{\mathrm{vr}}} \gamma_{t,k}^2 \|\boldsymbol{H}^{\mathrm{vr}}_{t,k}\|^2,$$

and with the convention that $\gamma_{t,0} := 0$, it holds

$$\mathbb{E}^{\mathcal{F}_{t,k}}\big[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\big] \le \|\boldsymbol{H}^{\mathrm{vr}}_{t,k} - \boldsymbol{h}(\boldsymbol{w}_{t,k-1})\|^2 + (L^2/\mathsf{b}_{\mathrm{vr}}) \gamma_{t,k}^2 \|\boldsymbol{H}^{\mathrm{vr}}_{t,k}\|^2.$$

When $\gamma_{t,k+1} \le \gamma_{t,k}$ for any $k \ge 1$, Lemma 15 implies that

$$\gamma_{t,k+1} \mathbb{E}^{\mathcal{F}_{t,0}}\big[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\big] \tag{134}$$

$$\le \frac{2L^2}{\mathsf{b}_{\mathrm{vr}}} \sum_{\ell=1}^{k} \gamma_{t,\ell}^3 \mathbb{E}^{\mathcal{F}_{t,0}}\big[\|\boldsymbol{H}^{\mathrm{vr}}_{t,\ell} - \boldsymbol{h}(\boldsymbol{w}_{t,\ell-1})\|^2 + \|\boldsymbol{h}(\boldsymbol{w}_{t,\ell-1})\|^2\big].$$

In the simple case when the stepsize sequence is constant ($\gamma_{t,k} = \gamma$) and the mean field is bounded ($c_{\boldsymbol{h},1} = 0$), the summed MSE along the inner loop iterations satisfies

$$\left(1 - 2\gamma^2 \frac{L^2 \mathrm{k_{in}}}{\mathsf{b}_{\mathrm{vr}}}\right) \sum_{k=0}^{\mathrm{k_{in}}-1} \mathbb{E}^{\mathcal{F}_{t,0}}\big[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\big] \le 2\gamma^2 (L^2 \mathrm{k_{in}}^2 / \mathsf{b}_{\mathrm{vr}}) c_{\boldsymbol{h},0}.$$

This inequality illustrates the benefit of variance reduction: the cumulated MSE can be set arbitrarily small by a convenient choice of the learning rate γ. Such a property remains true when the stepsize sequence is not constant and the mean field $\boldsymbol{h}$ is not bounded; it will be a key ingredient for the convergence analysis provided in Theorem 7 below.

The following lemma is an analogue of the Robbins-Siegmund Lemma for obtaining non-asymptotic bounds.

Lemma 16. *Assume **H** 1-b) and **H** 2. For any $t \in \{1, \dots, \mathrm{k_{out}}\}$ and $k \in \{0, \dots, \mathrm{k_{in}} - 1\}$, it holds*

$$\mathbb{E}^{\mathcal{F}_{t,0}}\big[\mathrm{V}(\boldsymbol{w}_{t,k+1})\big] \le \mathbb{E}^{\mathcal{F}_{t,0}}\big[\mathrm{V}(\boldsymbol{w}_{t,k})\big]$$

$$
\begin{aligned}
&- \gamma_{t,k+1}\{\nu_{t,k+1}\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathrm{W}(\boldsymbol{w}_{t,k})\right] \\
&\qquad + \mu_{t,k+1}\,\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right]\} \\
&+ \gamma_{t,k+1}\,\mathrm{a}\,\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right] + \gamma^2_{t,k+1} L_{\mathrm{V}} c_{\boldsymbol{h},0}.
\end{aligned}
$$

where $\nu_{t,k+1} := \varrho/2 - c_{\boldsymbol{h},1}\gamma_{t,k+1}L_{\mathrm{V}}$, $\mu_{t,k+1} := \varrho/2 - \gamma_{t,k+1}L_{\mathrm{V}}$ *and* $\mathrm{a} := (c_{\mathrm{V}}^2\varrho^{-1} + \varrho)/2$.

We will use (134) to show that the term $\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right]$ in the RHS is negligible *w.r.t.* the term $\mu_{t,k+1}\,\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right]$ in the LHS. We can establish a non-asymptotic convergence result.

Theorem 7. *Assume* ***H*** *1-b),* ***H*** *2 and* ***VR*** *1. Let* $\mathrm{k_{in}}$, $\mathrm{k_{out}}$ *be positive integers and* $\boldsymbol{w}_{\mathrm{init}} \in \mathbb{R}^d$. *Let* $\{\gamma_{t,k+1}, 1 \le t \le \mathrm{k_{out}}, 0 \le k \le \mathrm{k_{in}} - 1\}$ *be a stepsize sequence satisfying for all* $k \ge 1$: $\gamma_{t,k+1} \le \gamma_{t,k}$, $(1 \vee c_{\boldsymbol{h},1})\,\gamma_{t,k}\,\lambda_{t,k} \in (0, \varrho/2)$ *where* $\lambda_{t,k} := L_{\mathrm{V}} + \gamma_{t,k}L^2\left\{c_{\mathrm{V}}^2\varrho^{-1} + \varrho\right\}\mathrm{k_{in}}/\mathsf{b}_{\mathrm{vr}}$. *Consider the sequence given by Algorithm 1. Then*

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=1}^{\mathrm{k_{in}}}\gamma_{t,k}\left(\frac{\varrho}{2} - c_{\boldsymbol{h},1}\,\gamma_{t,k}\,\lambda_{t,k}\right)\mathbb{E}\left[\mathrm{W}(\boldsymbol{w}_{t,k-1})\right] \\
&+ \textstyle\sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=1}^{\mathrm{k_{in}}}\gamma_{t,k}\left(\frac{\varrho}{2} - \gamma_{t,k}\,\lambda_{t,k}\right)\mathbb{E}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k} - \boldsymbol{h}(\boldsymbol{w}_{t,k-1})\|^2\right] \\
&\le \mathbb{E}\left[\mathrm{V}(\boldsymbol{w}_{\mathrm{init}})\right] - \mathrm{V}_\star + c_{\boldsymbol{h},0}\,\mathrm{B^{vr}},
\end{aligned}
\tag{135}
$$

where

$$
\mathrm{B^{vr}} := L_{\mathrm{V}}\sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=1}^{\mathrm{k_{in}}}\gamma^2_{t,k} + \frac{L^2\mathrm{k_{in}}}{\mathsf{b}_{\mathrm{vr}}}\left(\frac{c_{\mathrm{V}}^2}{\varrho} + \varrho\right)\sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=1}^{\mathrm{k_{in}}}\gamma^3_{t,k}.
$$

Theorem 7 controls $\mathrm{W}(\boldsymbol{w}_{t,k})$ and the quadratic error $\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2$ along the $\mathrm{k_{in}}\mathrm{k_{out}}$ iterations. Set

$$
T := \mathrm{k_{in}}\mathrm{k_{out}}.
$$

We observe the following consequences of Theorem 7.

Random stopping. The LHS in Theorem 7 can be viewed as the expected value of $\mathrm{W}(\boldsymbol{w}_{R^{\mathrm{vr}}_T})$ and $\|\boldsymbol{H}^{\mathrm{vr}}_{\tilde{R}^{\mathrm{vr}}_T+1} - \boldsymbol{h}(\boldsymbol{w}_{\tilde{R}^{\mathrm{vr}}_T})\|^2$ where R^{vr}_T and $\tilde{R}^{\mathrm{vr}}_T$ are random variables taking values in $\{1,\ldots,\mathrm{k_{out}}\}\times\{0,\ldots,\mathrm{k_{in}}-1\}$, independent of the $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$, and with probability mass functions (see Section IV-A)

$$
\begin{aligned}
\mathbb{P}(R^{\mathrm{vr}}_T = (t,k)) &\propto \gamma_{k+1}\left(\varrho/2 - c_{\boldsymbol{h},1}\,\gamma_{t,k+1}\,\lambda_{t,k+1}\right) \\
\mathbb{P}(\tilde{R}^{\mathrm{vr}}_T = (t,k)) &\propto \gamma_{k+1}\left(\varrho/2 - \gamma_{t,k+1}\,\lambda_{t,k+1}\right).
\end{aligned}
$$

Remark 4. *Similar to Remark 3. When the function* W *is convex, a stronger convergence result can be derived. Define the convex combination of the iterates*

$$
\bar{\boldsymbol{w}}^{\mathrm{vr}}_T := \sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=0}^{\mathrm{k_{in}}-1}\frac{\gamma_{t,k+1}\,\omega^{\mathrm{vr}}_{t,k+1}}{\sum_{t'=1}^{\mathrm{k_{out}}}\sum_{k'=0}^{\mathrm{k_{in}}-1}\gamma_{t',k'+1}\,\omega^{\mathrm{vr}}_{t',k'+1}},
$$

where $\omega^{\mathrm{vr}}_{t,k+1} := \varrho/2 - c_{\boldsymbol{h},1}\,\gamma_{t,k+1}\,\lambda_{t,k+1}$. *We have from Theorem 7 that* $\mathbb{E}\left[\mathrm{W}(\bar{\boldsymbol{w}}^{\mathrm{vr}}_T)\right]$ *is upper bounded by the RHS of* (135) *divided by* $\sum_{t'=1}^{\mathrm{k_{out}}}\sum_{k'=0}^{\mathrm{k_{in}}-1}\gamma_{t',k'+1}\,\omega^{\mathrm{vr}}_{t',k'+1}$.

Constant stepsize. Assume $\gamma_{t,k+1} = \gamma$ for each $t \in \{1,\ldots,\mathrm{k_{out}}\}$ and $k \in \{0,\ldots,\mathrm{k_{in}}-1\}$. The assumptions of Theorem 7 are satisfied by choosing $\gamma \in (0, \gamma^{\mathrm{vr}}_{\max})$ where

$$
\gamma^{\mathrm{vr}}_{\max} := \frac{\varrho L_{\mathrm{V}}\mathsf{b}_{\mathrm{vr}}}{2L^2(c_{\mathrm{V}}^2 + \varrho^2)\mathrm{k_{in}}}\left(\left\{1 + 2\frac{L^2(c_{\mathrm{V}}^2 + \varrho^2)\mathrm{k_{in}}}{L_{\mathrm{V}}^2(1 \vee c_{\boldsymbol{h},1})\mathsf{b}_{\mathrm{vr}}}\right\}^{1/2} - 1\right).
$$

This yields

$$
\begin{aligned}
&\tfrac{1}{T}\textstyle\sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=0}^{\mathrm{k_{in}}-1}\left(\mathbb{E}[\mathrm{W}(\boldsymbol{w}_{t,k})] + \mathbb{E}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k}\|^2\right]\right) \\
&\le \frac{\Delta_1}{\gamma T\{\varrho/2 - \gamma\lambda(\gamma)(1 \vee c_{\boldsymbol{h},1})\}} \\
&\quad + \gamma\, c_{\boldsymbol{h},0}\left(\frac{L_{\mathrm{V}} + \gamma^{\mathrm{vr}}_{\max}L^2\left(c_{\mathrm{V}}^2 + \varrho^2\right)\mathrm{k_{in}}/(\varrho\mathsf{b}_{\mathrm{vr}})}{\varrho/2 - \gamma\lambda(\gamma)(1 \vee c_{\boldsymbol{h},1})}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda(\gamma) &:= L_{\mathrm{V}} + \gamma L^2\left\{c_{\mathrm{V}}^2\varrho^{-1} + \varrho\right\}\mathrm{k_{in}}/\mathsf{b}_{\mathrm{vr}}, \\
\Delta_1 &:= \mathbb{E}[\mathrm{V}(\boldsymbol{w}_{\mathrm{init}})] - \mathrm{V}_\star.
\end{aligned}
$$

Contrary to Theorem 1 (see Corollary 3), the RHS can be made small by a clever choice of γ: even if the oracle $\boldsymbol{H}^{\mathrm{vr}}_{t,k+1}$ is a biased estimator of $\boldsymbol{h}(\boldsymbol{w}_{t,k})$ (see Lemma 14), the SPIDER variance reduction is able to manage this bias.

The terms in the RHS can be adjusted according to the total number of iterations T and the different parameters of the problem by a suitable choice of γ. When $\gamma \le \gamma^{\mathrm{vr}}_{\max}/2$, then $\varrho/2 - \gamma\lambda(\gamma)(1 \vee c_{\boldsymbol{h},1}) \ge \varrho/4$; in addition, the function $\gamma \mapsto \alpha/\gamma + \beta\gamma$ is minimized on $(0, \gamma^{\mathrm{vr}}_{\max}/2]$ at the point $\sqrt{\alpha/\beta} \wedge (\gamma^{\mathrm{vr}}_{\max}/2)$ when $\alpha, \beta > 0$. Therefore, set

$$
\Delta_2 := 2L_{\mathrm{V}}\varrho + \gamma^{\mathrm{vr}}_{\max}L^2\left(c_{\mathrm{V}}^2 + \varrho^2\right)\mathrm{k_{in}}/\mathsf{b}_{\mathrm{vr}},
$$

and $\gamma^{\mathrm{vr}}_T := \gamma^{\mathrm{vr}}_{\max}/2$ if $c_{\boldsymbol{h},0} = 0$ and otherwise

$$
\gamma^{\mathrm{vr}}_T := \sqrt{2\Delta_1\varrho/T\Delta_2} \wedge (\gamma^{\mathrm{vr}}_{\max}/2). \tag{136}
$$

Corollary 10 (of Theorem 7)**.** *Setting* $\gamma_{t,k+1} = \gamma^{\mathrm{vr}}_T$ *for* $t \in \{1,\ldots,\mathrm{k_{out}}\}$ *and* $k \in \{0,\ldots,\mathrm{k_{in}}-1\}$ *we get*

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{\mathrm{k_{out}}}\sum_{k=0}^{\mathrm{k_{in}}-1}\left(\mathbb{E}[\mathrm{W}(\boldsymbol{w}_{t,k})] + \mathbb{E}\left[\|\boldsymbol{H}^{\mathrm{vr}}_{t,k+1} - \boldsymbol{h}(\boldsymbol{w}_{t,k}\|^2\right]\right) \\
&\qquad\qquad \le 4\Delta_1/(\gamma^{\mathrm{vr}}_T T\varrho) + 2\gamma^{\mathrm{vr}}_T\, c_{\boldsymbol{h},0}\Delta_2/\varrho^2\,.
\end{aligned}
$$

Choose $\mathrm{k_{in}} = \mathsf{b}_{\mathrm{vr}}$ and let us comment the rate of the RHS when $T \to +\infty$. When $c_{\boldsymbol{h},0} = 0$, the RHS decreases at the rate $O(1/T)$, while when $c_{\boldsymbol{h},0} > 0$, the rate is $O(1/\sqrt{T})$.

ϵ-Approximate Stationarity. Using Corollary 10, we analyze the complexity of SA-SPIDER to reach ϵ-approximate stationarity. First, observe that the RHS in Corollary 10 is an upper bound of $\mathbb{E}\left[\mathrm{W}(\boldsymbol{w}_{R^{\mathrm{vr}}_T})\right]$ where R^{vr}_T is a uniform random variable on $\{1,\ldots,\mathrm{k_{out}}\}\times\{0,\ldots,\mathrm{k_{in}}-1\}$.

Consider first the case when $c_{\boldsymbol{h},0} = 0$; then $\gamma^{\mathrm{vr}}_T = \gamma^{\mathrm{vr}}_{\max}/2$ and is independent of the accuracy ϵ. Choose $\mathrm{k_{out}} = \mathcal{O}(1/(\sqrt{n}\epsilon))$ and $\mathrm{k_{in}} = \mathsf{b}_{\mathrm{vr}} = \lceil\sqrt{n}\rceil$; it means that the total number of calls to one of the functions $\boldsymbol{h}_i$ (see (130)) is equal to n. Then the total number of iterations to reach an ϵ-approximate stationary point is $\mathcal{O}(1/\epsilon)$ and the total number of calls to one of the functions $\boldsymbol{h}_i$ is $n\mathrm{k_{out}} + 2\mathrm{k_{out}}\mathrm{k_{in}}\mathsf{b}_{\mathrm{vr}} = \mathcal{O}(\sqrt{n}/\epsilon)$. Such a complexity analysis retrieves earlier results established in specific settings of SA-SPIDER: SG for non convex optimization [41, Theorem 2], [42, Theorem 1], the stochastic EM algorithms [33, Section 3] and for more general SA-based root-finding problems [38, Corollary 4.3].

Consider now the case $c_{\boldsymbol{h},0} > 0$. Choose again $\mathrm{k_{in}} = \mathsf{b}_{\mathrm{vr}} = \lceil n \rceil$, and $\mathrm{k_{out}}$ large enough so that

$$T \geq \frac{8(1+c_{\boldsymbol{h},0})^2 \Delta_1 \Delta_2}{\varrho^3 \epsilon^2} \vee \frac{8\Delta_1 \varrho}{\Delta_2 (\gamma_{\max}^{\mathrm{vr}})^2}.$$

Then $\gamma_T^{\mathrm{vr}} = \sqrt{2\Delta_1 \varrho}/\sqrt{T\Delta_2}$ and SA-SPIDER reaches an ϵ-approximate stationary point in $T = \mathcal{O}(1/\epsilon^2)$ iterations, and by calling $\mathcal{O}(\sqrt{n}/\epsilon^2)$ functions $\boldsymbol{h}_i$. To our best knowledge, this is the first complexity analysis of SA-SPIDER when $c_{\boldsymbol{h},0} > 0$.

VII. Conclusions

We have overview state-of-the-art results for SA scheme with a focus on its use as a general stochastic non-gradient algorithm common to statistical learning. We have proposed a general theoretical framework based on the designs of flexible Lyapunov function and unified the modern asymptotic, non-asymptotic convergence results for SA schemes. We illustrated the applications of our techniques to SG, compressed SA, stochastic EM and TD learning; as well as presenting how the recent variance reduction technique can be adopted to SA. We studied the effects of *bias* in SA updates caused by the non-gradient nature in popular designs.

Our findings shed lights on how to design stochastic algorithms with nice convergence properties. In particular, we illustrated how to construct the stochastic random field in SA from fixed point equation of the statistical learning problem, and to tame with the bias in SA resulted from the design.

References

[1] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[2] J. R. Blum, "Multidimensional stochastic approximation methods," *The Annals of Mathematical Statistics*, pp. 737–744, 1954.

[3] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE transactions on automatic control*, vol. 22, no. 4, pp. 551–575, 1977.

[4] L. Ljung and T. Söderström, *Theory and practice of recursive identification*. MIT press, 1983.

[5] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 2012, vol. 22.

[6] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2002.

[7] L. Bottou, "Stochastic learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 146–168.

[8] ——, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[9] J. C. Príncipe, W. Liu, and S. Haykin, *Kernel adaptive filtering: a comprehensive introduction*. John Wiley & Sons, 2011.

[10] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[11] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 32–43, 2014.

[12] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

[13] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003, vol. 35.

[14] Y. Ren and D. Goldfarb, "Tensor normal training for deep learning models," *arXiv preprint arXiv:2106.02925*, 2021.

[15] S. Gunasekar, B. Woodworth, and N. Srebro, "Mirrorless mirror descent: A natural derivation of mirror descent," in *AISTATS*, 2021, pp. 2305–2313.

[16] A. Cichocki, R. Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electronics Letters*, vol. 30, no. 17, pp. 1386–1387, 1994.

[17] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, 1998.

[18] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[19] A. Cichocki and S.-i. Amari, *Adaptive blind signal and image processing: learning algorithms and applications*. John Wiley & Sons, 2002.

[20] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *NeurIPS*, 2015, pp. 3123–3131.

[21] A. Shekhovtsov and V. Yanush, "Reintroducing straight-through estimators as principled methods for stochastic binary networks," in *DAGM German Conference on Pattern Recognition*. Springer, 2021, pp. 111–126.

[22] J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke, "Coordinate descent converges faster with the gauss-southwell rule than random selection," in *ICML*. PMLR, 2015, pp. 1632–1641.

[23] K. Lange, *Optimization*, ser. Springer Texts in Statistics. Springer New York, 2013.

[24] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[25] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the em algorithm," *Annals of statistics*, pp. 94–128, 1999.

[26] G. Fort, P. Gach, and E. Moulines, "Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence," *Statistics and Computing*, vol. 31, no. 4, pp. 1–24, 2021.

[27] B. Karimi, B. Miasojedow, E. Moulines, and H.-T. Wai, "Non-asymptotic analysis of biased stochastic approximation scheme," in *Conference on Learning Theory*. PMLR, 2019, pp. 1944–1974.

[28] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, 1988.

[29] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[30] J. Tsitsiklis and B. Van Roy, "Analysis of temporal-diffference learning with function approximation," *NeurIPS*, vol. 9, 1996.

[31] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[32] J. Chen, J. Zhu, Y. W. Teh, and T. Zhang, "Stochastic expectation maximization with variance reduction," *NeurIPS*, vol. 31, 2018.

[33] G. Fort, E. Moulines, and H.-T. Wai, "A Stochastic Path-Integrated Differential Estimator Expectation Maximization Algorithm," in *NeurIPS*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[34] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[35] A. H. Sayed, *Adaptive filters*. John Wiley & Sons, 2011.

[36] S. Meyn, *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

[37] G. Fort, E. Moulines, A. Schreck, and M. Vihola, "Convergence of Markovian Stochastic Approximation with Discontinuous Dynamics," *SIAM Journal on Control and Optimization*, vol. 54, no. 2, pp. 866–893, 2016.

[38] G. Fort and E. Moulines, "Stochastic Variable Metric Proximal Gradient with variance reduction for non-convex composite optimization," *Statistics and Computing*, 2023.

[39] V. B. Tadić and A. Doucet, "Asymptotic bias of stochastic gradient search," *The Annals of Applied Probability*, vol. 27, no. 6, pp. 3255–3304, 2017.

[40] A. Ramaswamy and S. Bhatnagar, "Analysis of gradient descent methods with nondiminishing bounded errors," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1465–1471, 2017.

[41] C. Fang, C. J. Li, Z. Lin, and T. Zhang, "SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator," *NeurIPS*, vol. 31, 2018.

[42] Z. Wang, K. Ji, Y. Zhou, Y. Liang, and V. Tarokh, "Spiderboost and momentum: Faster variance reduction algorithms," in *NeurIPS*, 2019.

[43] A. Dieuleveut, G. Fort, E. Moulines, and G. Robin, "Federated-EM with heterogeneity mitigation and variance reduction," *NeurIPS*, vol. 34, 2021.

[44] W. Su, S. Boyd, and E. J. Candès, "A differential equation for modeling nesterov's accelerated gradient method: theory and insights," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5312–5354, 2016.

[45] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, "Understanding the acceleration phenomenon via high-resolution differential equations," *Mathematical Programming*, pp. 1–70, 2021.

[46] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "Robust accelerated gradient methods for smooth strongly convex functions," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 717–751, 2020.

[47] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," *NeurIPS*, vol. 20, 2007.

[48] A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar, "Information-theoretic lower bounds on the oracle complexity of convex optimization," *NeurIPS*, vol. 22, 2009.

[49] A. Nemirovski, "On parallel complexity of nonsmooth convex optimization," *Journal of Complexity*, vol. 10, no. 4, pp. 451–463, 1994.

[50] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[51] A. Juditsky and A. Nemirovski, "First order methods for nonsmooth convex large-scale optimization, i: general purpose methods," *Optimization for Machine Learning*, vol. 30, no. 9, pp. 121–148, 2011.

[52] A. Juditsky, A. Nemirovski *et al.*, "First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure," *Optimization for Machine Learning*, vol. 30, no. 9, pp. 149–183, 2011.

[53] E. Hazan and S. Kale, "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2489–2512, 2014.

[54] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016.

[55] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[56] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches." *Journal of Machine Learning Research*, vol. 13, no. 1, 2012.

[57] D. Needell, R. Ward, and N. Srebro, "Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm," *NeurIPS*, vol. 27, 2014.

[58] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth annual conference of the international speech communication association*, 2014.

[59] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding," *NeurIPS*, vol. 30, pp. 1709–1720, 2017.

[60] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[61] E. De Klerk, F. Glineur, and A. B. Taylor, "Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation," *SIAM Journal on Optimization*, vol. 30, no. 3, pp. 2053–2082, 2020.

[62] O. Gannot, "A frequency-domain analysis of inexact gradient methods," *Mathematical Programming*, vol. 194, no. 1, pp. 975–1016, 2022.

[63] C. M. De Sa, C. Zhang, K. Olukotun, and C. Ré, "Taming the wild: A unified analysis of hogwild-style algorithms," *NeurIPS*, vol. 28, 2015.

[64] S. Magnússon, C. Enyioha, N. Li, C. Fischione, and V. Tarokh, "Convergence of limited communication gradient methods," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1356–1371, 2017.

[65] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi, "Error feedback fixes signsgd and other gradient compression schemes," in *ICML*. PMLR, 2019, pp. 3252–3261.

[66] S. U. Stich and S. P. Karimireddy, "The error-feedback framework: Better rates for sgd with delayed gradients and compressed updates," *Journal of Machine Learning Research*, vol. 21, pp. 1–36, 2020.

[67] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signsgd: Compressed optimisation for non-convex problems," in *ICML*. PMLR, 2018, pp. 560–569.

[68] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[69] M. G. Rabbat and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 798–808, 2005.

[70] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks-part i: Gaussian case," *IEEE transactions on signal processing*, vol. 54, no. 3, pp. 1131–1143, 2006.

[71] E. J. Msechu and G. B. Giannakis, "Sensor-centric data reduction for estimation with wsns via censoring and quantization," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 400–414, 2011.

[72] P. Yi and Y. Hong, "Quantized subgradient algorithm and data-rate analysis for distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 4, pp. 380–392, 2014.

[73] H. Mania, X. Pan, D. Papailiopoulos, B. Recht, K. Ramchandran, and M. I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2202–2229, 2017.

[74] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright, "Randomized smoothing for stochastic optimization," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 674–701, 2012.

[75] K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee, "Optimal algorithms for non-smooth distributed optimization in networks," *NeurIPS*, vol. 31, 2018.

[76] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," *NeurIPS*, vol. 24, 2011.

[77] S. Chaturapruek, J. C. Duchi, and C. Ré, "Asynchronous stochastic convex optimization: the noise is in the noise and sgd don't care," *NeurIPS*, vol. 28, 2015.

[78] C. Philippenko and A. Dieuleveut, "Preserved central model for faster bidirectional compression in distributed settings," *NeurIPS*, vol. 34, pp. 2387–2399, 2021.

[79] M. Courbariaux, Y. Bengio, and J.-P. David, "Training deep neural networks with low precision multiplications," *arXiv preprint arXiv:1412.7024*, 2014.

[80] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep learning with limited numerical precision," in *ICML*. PMLR, 2015, pp. 1737–1746.

[81] C. De Sa, M. Leszczynski, J. Zhang, A. Marzoev, C. R. Aberger, K. Olukotun, and C. Ré, "High-accuracy low-precision training," *arXiv preprint arXiv:1803.03383*, 2018.

[82] H. Li, S. De, Z. Xu, C. Studer, H. Samet, and T. Goldstein, "Training quantized nets: A deeper understanding," *NeurIPS*, vol. 30, 2017.

[83] H. Le, R. K. Høier, C.-T. Lin, and C. Zach, "Adaste: An adaptive straight-through estimator to train binary neural networks," *arXiv preprint arXiv:2112.02880*, 2021.

[84] Z. Liu, K.-T. Cheng, D. Huang, E. P. Xing, and Z. Shen, "Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation," in *CVPR*, 2022, pp. 4942–4952.

[85] A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end feedback loss in speech chain framework via straight-through estimator," in *ICASSP*. IEEE, 2019, pp. 6281–6285.

[86] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc. B Met.*, vol. 39, no. 1, pp. 1–38, 1977.

[87] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[88] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[89] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM review*, vol. 26, no. 2, pp. 195–239, 1984.

[90] A. Swami, "Non-gaussian mixture models for detection and estimation in heavy-tailed noise," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 6. IEEE, 2000, pp. 3802–3805.

[91] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[92] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*. New York: Springer, 2005.

[93] Z. Ghahramani and M. Jordan, "Factorial hidden markov models," *NeurIPS*, vol. 8, 1995.

[94] B. Everitt, *An introduction to latent variable models*. Chapman and Hall London ; New York, 1984.

[95] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on signal processing*, vol. 42, no. 10, pp. 2664–2677, 1994.

[96] M. R. Gupta, Y. Chen *et al.*, "Theory and use of the em algorithm," *Foundations and Trends® in Signal Processing*, vol. 4, no. 3, pp. 223–296, 2011.

[97] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[98] L. Brown, *Fundamentals of statistical exponential families : with applications in statistical decision theory*, ser. Lecture notes-monograph

series Fundamentals of statistical exponential families. Institute of Mathematical Statistics, 1986.

[99] G. Celeux and J. Diebolt, "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quarterly*, vol. 2, pp. 73–82, 1985.

[100] G. Wei and M. Tanner, "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *J. Am. Stat. Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.

[101] G. Fort and E. Moulines, "Convergence of the Monte Carlo Expectation Maximization for curved exponential families," *Ann. Statist.*, vol. 31, no. 4, pp. 1220–1259, 2003.

[102] R. M. Neal and G. E. Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Dordrecht: Springer Netherlands, 1998, pp. 355–368.

[103] S. K. Ng and G. J. McLachlan, "On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures," *Stat. Comput.*, vol. 13, no. 1, pp. 45–55, 2003.

[104] B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle, "On the Global Convergence of (Fast) Incremental Expectation Maximization Methods," in *NeurIPS*, 2019, pp. 2837–2847.

[105] F. Kunstner, R. Kumar, and M. Schmidt, "Homeomorphic-invariance of em: Non-asymptotic convergence in kl divergence for exponential families via mirror descent," in *AISTATS*, 2021, pp. 3295–3303.

[106] F. Maire, S. Lefebvre, R. Douc, and E. Moulines, "An online learning algorithm for mixture models of deformable templates," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.

[107] H. D. Nguyen, F. Forbes, and G. J. Mclachlan, "Mini-batch learning of exponential family finite mixture models," *Statistics and Computing*, vol. 30, pp. 731–748, 2020.

[108] G. Oudoumanessah, M. Dojat, and F. Forbes, "Unsupervised scalable anomaly detection: application to medical imaging," hal-03824951, Research Report, 2022.

[109] J.-Y. Tourneret, M. Doisy, and M. Lavielle, "Bayesian off-line detection of multiple change-points corrupted by multiplicative noise: application to SAR image edge detection," *Signal Processing*, vol. 83, no. 9, pp. 1871–1887, 2003.

[110] M. Sahmoudi, K. Abed-Meraim, M. Lavielle, E. Kuhn, and P. Ciblat, "Blind source separation of noisy mixtures using a semi-parametric approach with application to heavy-tailed signals," in *EUSIPCO*, 2005.

[111] S. Allassonniere, E. Kuhn, A. Trouve, and Y. Amit, "Generative Model and Consistent Estimation Algorithms for Non-Rigid Deformable Models," in *ICASSP*, vol. 5, 2006.

[112] F. Septier, Y. Delignon, A. Menhaj-Rivenq, and C. Garnier, "Monte Carlo Methods for Channel, Phase Noise, and Frequency Offset Estimation With Unknown Noise Variances in OFDM Systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3613–3626, 2008.

[113] F. Richard, A. M. M. Samson, and C.-A. Cuenod, "A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences," *Stat. Comput.*, vol. 19, no. 4, pp. 465–478, 2009.

[114] S. Yildirim, A. T. Cemgil, and A. B. Ertuzun, "A hybrid method for deconvolution of bernoulli-gaussian processes," in *ICASSP*, 2009, pp. 3417–3420.

[115] F. Lindsten, "An efficient stochastic approximation EM algorithm using conditional particle filters," in *ICASSP*, 2013, pp. 6274–6278.

[116] M. Zhang, N. Singh, and P. T. Fletcher, "Bayesian Estimation of Regularization and Atlas Building in Diffeomorphic Image Registration," in *Information Processing in Medical Imaging*, J. C. Gee, S. Joshi, K. M. Pohl, W. M. Wells, and L. Zöllei, Eds. Springer Berlin Heidelberg, 2013, pp. 37–48.

[117] A. Svensson, T. B. Schön, and F. Lindsten, "Identification of jump Markov linear models using particle filters," in *53rd IEEE Conference on Decision and Control*, 2014, pp. 6504–6509.

[118] A. Boisbunon and J. Zerubia, "Estimation of the weight parameter with SAEM for marked point processes applied to object detection," in *EUSIPCO*, 2014, pp. 2185–2189.

[119] H. Braham, S. B. Jemaa, G. Fort, E. Moulines, and B. Sayrac, "Spatial Prediction Under Location Uncertainty in Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7633–7643, 2016.

[120] J. Liu, S. Kumar, and D. P. Palomar, "Parameter Estimation of Heavy-Tailed AR Model With Missing Data Via Stochastic EM," *IEEE Transactions on Signal Processing*, vol. 67, no. 8, pp. 2159–2172, 2019.

[121] ——, "Parameter Estimation of Heavy-Tailed AR(p) Model from Incomplete Data," in *EUSIPCO*, 2019, pp. 1–5.

[122] R. Zhou, J. Liu, S. Kumar, and D. P. Palomar, "Student's t var modeling with missing data via stochastic EM and Gibbs sampling," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6198–6211, 2020.

[123] ——, "Parameter estimation for student'st var model with missing data," in *ICASSP*. IEEE, 2021, pp. 5145–5149.

[124] S. Yıldırım, M. Khalafi, T. Güzel, H. Satık, and M. Yılmaz, "Supply curves in electricity markets: A framework for dynamic modeling and monte carlo forecasting," *IEEE Transactions on Power Systems*, pp. 1–14, 2022.

[125] B. Sauty and S. Durrleman, "Riemannian Metric Learning for Progression Modeling of Longitudinal Datasets," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5.

[126] S. Asmussen and P. Glynn, *Stochastic Simulation: Algorithms and Analysis*, ser. Stochastic Modelling and Applied Probability. Springer New York, 2007.

[127] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, "Importance Sampling: Intrinsic Dimension and Computational Cost," *Statistical Science*, vol. 32, no. 3, pp. 405–431, 2017.

[128] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*. PMLR, 2018, pp. 1691–1692.

[129] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012, vol. 1.

[130] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[131] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *ICML*, 2018, pp. 1587–1596.

[132] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions On Automatic Control*, vol. 42, no. 5, 1997.

[133] S. Di-Castro, S. Mannor, and D. Di Castro, "Analysis of stochastic processes through replay buffers," in *ICML*, 2022, pp. 5039–5060.

[134] M. Métivier and P. Priouret, "Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant," *Probability Theory and related fields*, vol. 74, no. 3, pp. 403–428, 1987.

[135] C. Andrieu, É. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM Journal on control and optimization*, vol. 44, no. 1, pp. 283–312, 2005.

[136] C. Andrieu and M. Vihola, "Markovian stochastic approximation with expanding projections," *Bernoulli*, vol. 20, no. 2, pp. 545–585, 2014.

[137] G. Wang and G. B. Giannakis, "Finite-time error bounds for biased stochastic approximation with applications to q-learning," in *AISTATS*. PMLR, 2020, pp. 3015–3024.

[138] A. Ramaswamy and S. Bhatnagar, "Stability of stochastic approximations with "controlled markov" noise and temporal difference learning," *IEEE Transactions on Automatic Control*, vol. 64, no. 6, pp. 2614–2620, 2018.

[139] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, ser. CMS Books in Mathematics. Springer New York, 2011.

[140] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified sgd with memory," *NeurIPS*, vol. 31, 2018.

[141] M. Safaryan, E. Shulgin, and P. Richtárik, "Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor," *Information and Inference: A Journal of the IMA*, vol. 11, no. 2, pp. 557–580, 2022.

[142] S. Horvóth, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtárik, "Natural compression for distributed deep learning," in *Mathematical and Scientific Machine Learning*. PMLR, 2022, pp. 129–141.

[143] Z. Li and C. M. De Sa, "Dimension-free bounds for low-precision training," in *NeurIPS*, vol. 32, 2019.

[144] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *ICML*. PMLR, 2019, pp. 6155–6165.

[145] Y. Atchadé, G. Fort, and E. Moulines, "On Perturbed Proximal Gradient Algorithms," *JMLR*, vol. 18, no. 10, pp. 1–33, 2017.

[146] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov Chains*, ser. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2018.

[147] H. Robbins and D. Siegmund, "A convergence theorem for non negative almost supermartingales and some applications," in *Optimizing methods in statistics*. Elsevier, 1971, pp. 233–257.

[148] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM journal on control and optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[149] D. Ruppert, "Efficient estimations from a slowly convergent robbins-monro process," Cornell University Operations Research and Industrial Engineering, Tech. Rep., 1988.

[150] S. Wright, J. Nocedal *et al.*, "Numerical optimization," *Springer Science*, vol. 35, no. 67-68, p. 7, 1999.

[151] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, ""convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions," in *ICML*. PMLR, 2017, pp. 654–663.

[152] L. Lei, C. Ju, J. Chen, and M. I. Jordan, "Non-convex finite-sum optimization via scsg methods," *NeurIPS*, vol. 30, 2017.

[153] P. Jain, D. Nagaraj, and P. Netrapalli, "Making the last iterate of sgd information theoretically optimal," in *Conference on Learning Theory*. PMLR, 2019, pp. 1752–1755.

[154] P. Jain, D. M. Nagaraj, and P. Netrapalli, "Making the last iterate of sgd information theoretically optimal," *SIAM Journal on Optimization*, vol. 31, no. 2, pp. 1108–1130, 2021.

[155] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *ICML*. PMLR, 2013, pp. 71–79.

[156] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth, "Lower bounds for non-convex stochastic optimization," *Mathematical Programming*, pp. 1–50, 2022.

[157] E. Moulines and F. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," *NeurIPS*, 2011.

[158] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[159] A. S. Nemirovskij and D. B. Yudin, "Problem complexity and method efficiency in optimization," 1983.

[160] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Stochastic convex optimization." in *COLT*, vol. 2, no. 4, 2009, p. 5.

[161] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[162] E. De Klerk, F. Glineur, and A. B. Taylor, "Worst-case convergence analysis of inexact gradient and newton methods through semidefinite programming performance estimation," *SIAM Journal on Optimization*, vol. 30, no. 3, pp. 2053–2082, 2020.

[163] E. Gorbunov, D. Kovalev, D. Makarenko, and P. Richtárik, "Linearly converging error compensated sgd," *NeurIPS*, vol. 33, pp. 20 889–20 900, 2020.

[164] M. Benaïm, "Dynamics of stochastic approximation algorithms," in *Seminaire de probabilites XXXIII*. Springer, 1999, pp. 1–68.

[165] M. Benaim, "A dynamical system approach to stochastic approximations," *SIAM Journal on Control and Optimization*, vol. 34, no. 2, pp. 437–472, 1996.

[166] V. B. Tadić and A. Doucet, "Asymptotic bias of stochastic gradient search," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, 2011, pp. 722–727.

[167] S. Andradóttir, "A stochastic approximation algorithm with varying bounds," *Operations Research*, vol. 43, no. 6, pp. 1037–1048, 1995.

[168] H. Chen and Y. Zhu, "Stochastic approximation procedures with randomly varying truncations," *Science in China, Ser. A*, 1986.

[169] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.

[170] ——, "Stochastic approximations and differential inclusions, part ii: Applications," *Mathematics of Operations Research*, vol. 31, no. 4, pp. 673–695, 2006.

[171] ——, "Perturbations of set-valued dynamical systems, with applications to game theory," *Dynamic Games and Applications*, vol. 2, no. 2, pp. 195–205, 2012.

[172] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

[173] S. Majewski, B. Miasojedow, and E. Moulines, "Analysis of nonsmooth stochastic approximation: the differential inclusion approach," *arXiv preprint arXiv:1805.01916*, 2018.

[174] J.-P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory*. Springer Science & Business Media, 2012, vol. 264.

[175] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth analysis and control theory*. Springer Science & Business Media, 2008, vol. 178.

[176] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik, "Variance-reduced methods for machine learning," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 1968–1983, 2020.

[177] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *NeurIPS*, vol. 26, 2013.

[178] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," *NeurIPS*, vol. 27, 2014.

[179] X. Wang, S. Ma, D. Goldfarb, and W. Liu, "Stochastic Quasi-Newton Methods for Nonconvex Stochastic Optimization," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 927–956, 2017.

[180] L. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *ICML*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70, 2017, pp. 2613–2621.

[181] F. Shang, K. Zhou, H. Liu, J. Cheng, I. W. Tsang, L. Zhang, D. Tao, and L. Jiao, "VR-SGD: A Simple Stochastic Variance Reduction Method for Machine Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 1, pp. 188–202, 2020.

[182] Q. Zhang, F. Huang, C. Deng, and H. Huang, "Faster stochastic quasi-newton methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4388–4397, 2022.

[183] A. Han and J. Gao, "Improved Variance Reduction Methods for Riemannian Non-Convex Optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7610–7623, 2022.

[184] Y. Luo, X. Huo, and Y. Mei, "Implicit Regularization Properties of Variance Reduced Stochastic Mirror Descent," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 696–701.

[185] R. Twyman, S. Arridge, Z. Kereta, B. Jin, L. Brusaferri, S. Ahn, C. W. Stearns, B. F. Hutton, I. A. Burger, F. Kotasidis, and K. Thielemans, "An Investigation of Stochastic Variance Reduction Algorithms for Relative Difference Penalized 3D PET Image Reconstruction," *IEEE Transactions on Medical Imaging*, vol. 42, no. 1, pp. 29–41, 2023.

[186] G. Fort, E. Moulines, and H.-T. Wai, "Geom-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization," in *ICASSP*. IEEE, 2021, pp. 3135–3139.

[187] G. Fort and E. Moulines, "The Perturbed Prox-Preconditioned Spider Algorithm: Non-Asymptotic Convergence Bounds," in *IEEE SSP*, 2021, pp. 96–100.

[188] ——, "The Perturbed Prox-Preconditioned Spider Algorithm for EM-Based Large Scale Learning," in *IEEE SSP*, 2021, pp. 316–320.

[189] J. Neveu and T. Speed, *Discrete-parameter Martingales*, ser. Mathematical Studies. North-Holland, 1975.

[190] P. Hall and C. C. Heyde, *Martingale limit theory and its application*. Academic press, 2014.

Aymeric Dieuleveut received his Ph.D. degree in mathematical statistics from École Normale Supérieure de Paris, France, in 2017, after obtaining his MsC degree in mathematics from École Normale Supérieure de Paris in 2014. From 2017 to 2019, he was a Research Scientist at École Polytechnique Fédérale de Lausanne. In 2019, he became Assistant Professor at École Polytechnique, Paris, France.

His research topics cover high dimensional statistics, stochastic optimization, statistical machine learning and Federated Learning.

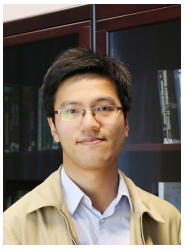
Gersende Fort received her Ph.D. degree in applied mathematics from University Paris VI - France in 2001, the Engineering degree from Ecole Nationale Supérieure des Télécommunications - France in 1997, and the Habilitation à Diriger les Recherches in 2010. In 2001, she joined the Centre National de la Recherche Scientifique (CNRS); she is now CNRS Senior Researcher at the Institut de Mathématiques de Toulouse - France.

Her areas of expertise cover statistical signal processing, computational statistics, statistical learning, stochastic optimization, Monte Carlo methods, with applications in inverse Bayesian problems. She has been an invited speaker at numerous conferences in computational statistics and machine learning.

Eric Moulines received his Ph. D. degree in electrical engineering from Ecole Nationale Supérieure des Télécommunications, in 1990, the Engineering degree from Ecole Polytechnique, Paris, France, in 1984. In 1990, he joined the Signal and Image processing department at Télécom Paris where he became a full professor in 1996. In 2015, he joined the Applied Mathematics Center of Ecole Polytechnique, where he is currently professor.

His areas of expertise include statistical signal processing, computational statistics (Monte Carlo simulations, stochastic approximation), statistical machine learning, and time-series analysis.

His current research topics cover high-dimensional Monte Carlo sampling, stochastic optimization, and generative models (variational autoencoders, Generative Adversarial Networks). He received the silver medal from the Centre National Recherche Scientifique (CNRS) in 2010, the "Grand Prix Orange" de l'Académie des Sciences (2011), the EURASIP technical Achievement Award (2020). He has been elected to the French Academy of Sciences in 2017, where he is currently deputy chair for computer sciences and applied mathematics. He is a fellow of the Institute of Mathematical Statistics (IMS) - 2016, and of the EURASIP - 2011.



Hoi-To Wai (S'11–M'18) received his Ph.D. degree from Arizona State University in Electrical Engineering in Fall 2017, B. Eng. (with First Class Honor) and M. Phil. degrees in Electronic Engineering from The Chinese University of Hong Kong (CUHK) in 2010 and 2012, respectively. He is currently an Assistant Professor in the Department of SEEM at CUHK (HK).

He currently serves on the editorial board of the IEEE Transactions on Signal and Information Processing over Networks. His research interests are in the broad area of optimization algorithms, graph signal processing, and machine learning. He has received a Best Student Paper Award from ICASSP 2018, and the 2017's Dean's Dissertation Award from the Ira A. Fulton Schools of Engineering of ASU for his thesis on network science and distributed optimization. His works have received a best student paper award from ICASSP 2018 and a best dissertation award from Arizona State University's Schools of Engineering.

Supplementary Material for "Stochastic Approximation Beyond Gradient for Signal Processing and Machine Learning"

Aymeric Dieuleveut, Gersende Fort, Eric Moulines, Hoi-To Wai

APPENDIX

This document presents the proofs or detailed calculations that have been skipped in the main paper. Readers can find the following content: Appendix B shows two elementary inequalities for numerical sequences. Appendix C and Appendix D check the assumptions for the examples on compressed SA and TD(0) learning. Appendix E gives a variant of Lemma 13 for the finite-time bounds of SA. Appendix F and Appendix G show how to obtain finite-time bounds for compressed SG methods and SAEM algorithms. Appendix H gives missing proof for the asymptotic convergence. Appendix I provides proofs for the variance reduced SA algorithm.

A. Table of notations

In the following, we provide complete notation tables, that aggregate the notations of the paper, with references to the points where each notation is introduced. Tables V to VIII respectively aggregate all notations used in the application to the four examples, namely SGD, compressed SA, EM, and TD learning.

TABLE V: Summary of notations used in the analysis of SGD, in Sections II-B1, III-B1 and IV-B1.

Notation	Object	Def. in
		Section II-B1
F	function to be minimized	
$\nabla F(\cdot)$	gradient of F and mean-field $\boldsymbol{h}$	
n	number of observations	
$(Z_1, \ldots, Z_n)$	observations	
ρ	distribution of $\boldsymbol{X}_{k+1}$	
ℓ	loss	
b	batch size	
$f_i := \ell(\boldsymbol{w}, Z_i)$	loss function on obs. i	
		Section III-B1
$L_{\nabla f_i}$	Smoothness of f_i	**SG**1.*a*)
$L_{\nabla f}$	Smoothness of f	
M	Uniform bound on $\|\nabla f_i(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|$	**SG**1.*b*)
$\mu > 0$	strong-convexity modulus of F	**CVX**2

AD and EM are with Ecole Polytechnique, CMAP, UMR 7641, France. GF is with Institut de Mathématiques de Toulouse, UMR5216, Université de Toulouse, CNRS; UPS, F-31062 Toulouse Cedex 9, France. HTW is with CUHK, Hong Kong. E-mails: aymeric.dieuleveut@polytechnique.edu, gersende.fort@math.univ-toulouse.fr, eric.moulines@polytechnique.edu, htwai@se.cuhk.edu.hk. Work partly supported by the *Fondation Simone et Cino Del Duca, Institut de France*, ANR under the program MaSDOL-19-CE23-0017-01, ANR-19-CHIA-SCAI-002, HKRGC Project 24203520, Hi!Paris *FLAG* project, and been carried out under the auspices of Lagrange research Center for Mathematics and Calculus.

TABLE VI: Summary of notations used in the analysis of Compressed and modified SA, in Sections II-B2, III-B2 and IV-B2.

Notation	Object	Def. in
		Section II-B2
F	function to be minimized	
$j_{k+1} \in \{1, \ldots, d\}$	chosen coordinate in the k-th iteration	eq. (12)
$\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$	the canonical basis of $\mathbb{R}^d$	
$\nabla_j F$	j-th coordinate of the gradient	
$\boldsymbol{\mathcal{C}} : \mathbb{R}^d \times \mathsf{U} \to \mathbb{R}^d$	Compression operator	
U	general state space	
μ_U	distribution of U	
$\mathrm{Rand}_\mathrm{h}, \mathrm{Top}_\mathrm{h}$	Ex. of sparsification-based compression operator	Remark 2.1
h	Sparsification parameter for $\mathrm{Rand}_\mathrm{h}, \mathrm{Top}_\mathrm{h}$	Remark 2.1
Q_d, Q_s	Ex. of quantization-based compression operator	Remark 2.2
Δ	quantization resolution	Remark 2.2
		Section III-B2
$(1 - \delta_{\mathcal{C}})$	Contractivness of $\mathcal{C}$	**CSA**1, eq. (55)
$\zeta_1, \zeta_2 \in \bar{\mathbb{R}}_+$	technical constants (from Young inequality)	Lemma 1
$\omega_{\mathcal{C}}$	Relative bound on variance of $\mathcal{C}$	**CSA**2, eq. (58)
$\kappa_{\mathcal{C}}$	Uniform bound on variance of $\mathcal{C}$	**CSA**3, eq. (60)

TABLE VII: Summary of notations used in the analysis of EM, in Sections II-B3, III-B3 and IV-B3.

Notation		Object	Def. in
			Section II-B3
$F(\boldsymbol{\theta}) := -\log \int_{\bar{\mathcal{Z}}} p(z; \boldsymbol{\theta}) \bar{\mu}(\mathrm{d}z)$		Intractable objective function (to be minimized)	eq. (20)
$\boldsymbol{\theta}$		parameter of the original optimization problem	
$p(z; \boldsymbol{\theta}) = \prod_{i=1}^n p_i(z_i; \boldsymbol{\theta})$		Product form; $z \in \bar{\mathcal{Z}}$ and $z_i \in \mathcal{Z}$	
μ		sigma-finite measure on the measurable set $\mathcal{Z}$	
$\pi_i(z_i; \boldsymbol{\theta})$	:=	A distribution on $\mathcal{Z}$	
$p_i(z_i; \boldsymbol{\theta}) / \int_{\mathcal{Z}} p_i(u; \boldsymbol{\theta}) \mu(\mathrm{d}u)$			
$\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}'}$		EM surrogate function tangent at $\boldsymbol{\theta}'$	eq. (22), (23)
		– In the maximum likelihood context –	
Y_i		observation $\#i$	
z_i		latent variable $\#i$	
$p_i(z_i; \boldsymbol{\theta})$		joint probability of the observation Y_i and the latent variable z_i for a given value of the parameter $\boldsymbol{\theta}$;	
$g_i(\boldsymbol{\theta}) := \int_{\mathcal{Z}} p_i(z_i; \boldsymbol{\theta}) \mu(\mathrm{d}z_i)$		likelihood of Y_i	eq. (21)
$\pi_i(z_i; \boldsymbol{\theta})$		posterior distribution of the latent variable z_i given the observation $\#i$ when the value of the parameter is $\boldsymbol{\theta}$	eq. (24)
		In the exponential family context	
S_i		sufficient statistics associated to the observation $\#i$	**EM**1
$\bar{\mathsf{s}}_i(\boldsymbol{\theta}) := \int_{\mathcal{Z}} \mathsf{S}_i(z_i) \pi_i(z_i; \boldsymbol{\theta}) \mu(\mathrm{d}z_i)$		expectation of the sufficient statistic	eq. (29)
$\bar{\mathsf{s}}(\boldsymbol{\theta}) := n^{-1} \sum_{i=1}^n \bar{\mathsf{s}}_i(\boldsymbol{\theta})$		the mean value of the n functions $\bar{\mathsf{s}}_i$	
$\mathcal{Q}^{\mathrm{EM}}_{\boldsymbol{\theta}'}(\cdot) := \langle \bar{\mathsf{s}}(\boldsymbol{\theta}') \mid \boldsymbol{\phi}(\cdot) \rangle - \boldsymbol{\psi}(\cdot)$		the specific form of the EM surrogate function	
$\boldsymbol{\phi}, \boldsymbol{\psi}$		functions on $\mathbb{R}^d$, parameterizing the family of surrogate functions	**EM**1
T		optimization map in EM	**EM**2
b_{EM}		batch size in Mini-batch EM	eq. (32)
m		number of Monte Carlo samples in SAEM	eq. (33), (34)
			Section III-B3
$F_\star$		Uniform lower bound on F	**EM**3
$\mathsf{B}(\boldsymbol{w})$		$d \times d$ p.d. matrix, s.t. $\nabla \mathrm{V}(\boldsymbol{w}) = -\mathsf{B}(\boldsymbol{w}) \boldsymbol{h}(\boldsymbol{w})$	**EM**3
$v_{\min} \le v_{\max}$		constants characterizing the conditioning of B	**EM**3
$\bar{\sigma}_0^2, \bar{\sigma}_1^2 \in \mathbb{R}_+$		control on the averaged sufficient statistic	**EM**4 and 5
$s_\star$		Uniform bound on $\|\mathsf{S}_i(z)\|$	**EM**6

TABLE VIII: Summary of notations used in the analysis of TD learning, in Sections II-B4, III-B4 and IV-B4.

Notation	Object	Def. in
		Section II-B4
π	policy in a Markov Decision Process	
$(\mathcal{S}, \mathcal{P}, \mathcal{R}, \lambda)$	Markov Reward Process (MRP)	
$\mathcal{S} = \{s_1, \dots, s_n\}$	state-space	
$\mathcal{P}$	$n \times n$ state transition matrix of the probability of transition	
$\mathcal{R}(s, s')$	reward function	
$\lambda \in (0, 1)$	discount factor	
$\mathcal{R}(s)$	expected instantaneous reward from state s	
ϖ	unique stationary distribution	**TD**1
$\mathcal{V}$	value function of the MRP	eq. (35)
$\{S_k,\ k \in \mathbb{N}\}$	Markov chain started at $S_0 = s$, with Markov kernel $\mathcal{P}$	
$\mathcal{V}_{\boldsymbol{w}}(s) := \boldsymbol{\phi}(s)^\top \boldsymbol{w}$	linear approximation of $\mathcal{V}$	
$\boldsymbol{\phi}(s) \in \mathbb{R}^d$	*feature vector* for the state $s \in \mathcal{S}$	
$\boldsymbol{w} \in \mathbb{R}^d$	parameter vector to be estimated	
$\boldsymbol{\Phi}$	$n \times d$ feature matrix	
		Section III-B4
$\lambda \in (0, 1)$	contraction modulus	Lemma 9
$\mathcal{V}_*$	unique value function in $\mathrm{span}(\boldsymbol{\Phi})$ which solves the fixed point of the projected Bellman eq.	Lemma 9
$v_{\min}$	minimal eigenvalue of $\boldsymbol{\Phi}^\top \mathbf{D}_\varpi \boldsymbol{\Phi}$	Lemma 12
$\boldsymbol{\Sigma}_\varpi$	feature covariance matrix	Equation (72)

B. Elementary inequalities

Lemma 17. *Let $a > 0$ and $\{\gamma_k,\ k \in \mathbb{N}\}$ be a sequence such that $\gamma_k < 1/a$ for any $k \geq 1$. Then, for any integer $k \geq 1$,*

$$\sum_{j=1}^{k+1} \gamma_j \prod_{l=j+1}^{k+1} (1 - \gamma_l a) = \frac{1}{a} \left\{ 1 - \prod_{l=1}^{k+1} (1 - \gamma_l a) \right\} \tag{137}$$

If in addition, for some $0 < b < a$ and all $k \geq 1$, $\gamma_k / \gamma_{k+1} \leq 1 + b\gamma_{k+1}$, then

$$\sum_{j=1}^{k+1} \gamma_j^2 \prod_{l=j+1}^{k+1} (1 - \gamma_l a) \leq \gamma_{k+1} / (a - b)\,. \tag{138}$$

Proof. Consider first (137). Let us denote $u_{j:k+1} := \prod_{l=j}^{k+1} (1 - \gamma_l a)$. Then, for $j \in \{1, \dots, k+1\}$, $u_{j+1:k+1} - u_{j:k+1} = a\gamma_j u_{j+1:k+1}$. Hence,

$$\begin{aligned}
\sum_{j=1}^{k+1} \gamma_j \prod_{l=j+1}^{k+1} (1 - \gamma_l a) &= \gamma_{k+1} + \sum_{j=1}^{k} \gamma_j u_{j+1:k+1} \\
&= \gamma_{k+1} + \frac{1}{a} \sum_{j=1}^{k} (u_{j+1:k+1} - u_{j:k+1}) \\
&= a^{-1} (1 - u_{1:k+1})\,.
\end{aligned} \tag{139}$$

Consider now (138).

$$\begin{aligned}
\sum_{j=1}^{k+1} \gamma_j^2 \prod_{l=j+1}^{k+1} (1 - \gamma_l a) &= \gamma_{k+1} \sum_{j=1}^{k+1} \frac{\gamma_j}{\gamma_{k+1}} \gamma_j \prod_{l=j+1}^{k+1} (1 - \gamma_l a) \\
&= \gamma_{k+1} \sum_{j=1}^{k+1} \gamma_j \prod_{l=j+1}^{k+1} \frac{\gamma_{l-1}}{\gamma_l} (1 - \gamma_l a)
\end{aligned}$$

where in the last equality, we used

$$\frac{\gamma_j}{\gamma_{k+1}} = \frac{\gamma_k}{\gamma_{k+1}} \frac{\gamma_k - 1}{\gamma_k} \cdots \frac{\gamma_j}{\gamma_{j+1}}.$$

Note that, since $\gamma_{l-1}/\gamma_l \leq 1 + b\gamma_l$ for $l \geq 2$, we have

$$\frac{\gamma_{l-1}}{\gamma_l} (1 - \gamma_l a) \leq (1 + b\gamma_l)(1 - \gamma_l a) \leq 1 - (a - b)\gamma_l.$$

Substituting into the above inequality yields

$$\sum_{j=1}^{k+1} \gamma_j^2 \prod_{l=j+1}^{k+1} (1 - \gamma_l a) \leq \gamma_{k+1} \sum_{j=1}^{k+1} \gamma_j \prod_{l=j+1}^{k+1} (1 - \gamma_l (a - b))\,.$$

Applying (139) implies $\sum_{j=1}^{k+1} \gamma_j^2 \prod_{l=j+1}^{k+1} (1 - \gamma_l a) \leq \gamma_{k+1} \frac{1}{a-b}$ and thus the lemma. □

C. Proofs of Section III-B2

Throughout the proof, we will use the shorthand notations

$$\begin{aligned}
\mathbf{Y}_{k+1} &:= \boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}), \\
\mathbf{Z}_{k+1} &:= \boldsymbol{\mathcal{C}}(\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1}), \mathbf{U}_{k+1}).
\end{aligned}$$

Observe that by **H** 1 for the oracle $\mathbf{Y}_{k+1}$,

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 \leq \tau_{0,k} + \tau_{1,k} \mathrm{W}(\boldsymbol{w}_k). \tag{140}$$

For the proofs of Lemma 1 and Lemma 2, define the filtrations $\{\mathcal{F}_{k+1/2}, k \in \mathbb{N}\}$ by

$$\mathcal{F}_{k+1/2} := \sigma\left(\boldsymbol{w}_0, \boldsymbol{X}_1, \mathbf{U}_1, \cdots, \boldsymbol{X}_k, \mathbf{U}_k, \boldsymbol{X}_{k+1}\right). \tag{141}$$

CSA1 and **CSA**2 claim

$$\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right] \leq (1 - \delta_{\boldsymbol{\mathcal{C}}}) \|\mathbf{Y}_{k+1}\|^2, \tag{142}$$

and **CSA**2

$$\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{Z}_{k+1}\right] = \mathbf{Y}_{k+1}. \tag{143}$$

For the proof of Lemma 3, we define the filtrations $\{\mathcal{F}_{k+1/2}, k \in \mathbb{N}\}$ by

$$\mathcal{F}_{k+1/2} := \sigma\left(\boldsymbol{w}_0, \mathbf{U}_1, \boldsymbol{X}_1, \cdots, \mathbf{U}_k, \boldsymbol{X}_k, \mathbf{U}_{k+1}\right). \tag{144}$$

Proof of Lemma 1. Let $k \geq 0$. We first prove **H** 1-*c)* for the compressed oracle $\mathbf{Z}_{k+1}$. We write, for any $\zeta_1 > 0$

$$\begin{aligned}
\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 &\leq (1 + \zeta_1) \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\right]\|^2 \\
&\quad + (1 + \zeta_1^{-1}) \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2.
\end{aligned}$$

The second term is upper bounded by (140). For the first one, by convexity of $\|\cdot\|^2$ and (142), it holds that

$$\begin{aligned}
&\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\right]\|^2 \leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right] \\
&\leq \mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right]\right] \leq (1 - \delta_{\boldsymbol{\mathcal{C}}})\ \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|^2\right].
\end{aligned}$$

We further observe that, by definition of the conditional expectation

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|^2\right] &= \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2 \\
&\quad + \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2\right] \\
&\leq (1 + \zeta_2) \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 \\
&\quad + (1 + \zeta_2^{-1}) \|\boldsymbol{h}(\boldsymbol{w}_k)\|^2 \\
&\quad + \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2\right],
\end{aligned}$$

for any $\zeta_2 > 0$. This yields, by using **H** 1,

$$\begin{aligned}
&\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|^2\right] \leq (1 + \zeta_2)\left(\tau_{0,k} + \tau_{1,k} \mathrm{W}(\boldsymbol{w}_k)\right) \\
&+ (1 + \zeta_2^{-1})\left(c_{h,0} + c_{h,1} \mathrm{W}(\boldsymbol{w}_k)\right) + \sigma_0^2 + \sigma_1^2 \mathrm{W}(\boldsymbol{w}_k).
\end{aligned} \tag{145}$$

Hence, **H** 1-*c)* holds for $\mathbf{Z}_{k+1}$ with the constants

$$\begin{aligned}
&\tau_{\ell,k;\boldsymbol{\mathcal{C}}} := ((1 + \zeta_1) + (1 + \zeta_2)(1 + \zeta_1^{-1})(1 - \delta_{\boldsymbol{\mathcal{C}}}))\tau_{\ell,k} + \\
&(1 + \zeta_2^{-1})(1 + \zeta_1^{-1})(1 - \delta_{\boldsymbol{\mathcal{C}}}) c_{h,\ell} + (1 + \zeta_1^{-1})(1 - \delta_{\boldsymbol{\mathcal{C}}}) \sigma_\ell^2.
\end{aligned}$$

We now prove **H** 1-*d)* for the compressed oracle $\mathbf{Z}_{k+1}$. Using again the convexity of $\|\cdot\|^2$, it holds

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right]\|^2\right]$$

$$= \mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{Z}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right]\|^2\right]\right]$$
$$\leq \mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right]\right]$$

using the fact that $\mathbf{Y}_{k+1}$ is $\mathcal{F}_{k+1/2}$ measurable, and the fact that for any real random variable U, we have $\mathbb{E}[U] = \arg\min_{c\in\mathbb{R}} \mathbb{E}[\|U-c\|^2]$.

By (142), we write

$$\mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right]\right] \leq (1-\delta_{\boldsymbol{\mathcal{C}}})\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|^2\right];$$

the expectation in the RHS is upper bounded by (145). which yields the bound (57). □

Proof of Lemma 2. We follow the same lines as in the proof of Lemma 1 except that we use (143). Let $k \geq 0$. We first prove **H** 1-*c)* for the compressed oracle $\mathbf{Z}_{k+1}$. We write

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 = \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{Z}_{k+1}\right]\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2;$$

with (143) and **H** 1 for $\mathbf{Y}_{k+1}$, this yields

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 \leq \tau_{0,k} + \tau_{1,k}\mathrm{W}(\boldsymbol{w}_k).$$

Hence we have $\tau_{\ell,k;\boldsymbol{\mathcal{C}}} := \tau_{\ell,k}$ for $\ell \in \{0,1\}$.

We now verify **H** 1-*d)* for $\mathbf{Z}_{k+1}$. Using (143), we write

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right]\|^2\right]$$
$$= \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2\right]$$
$$= \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right] + \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2\right]$$
$$+ 2\mathbb{E}^{\mathcal{F}_k}\left[\langle \mathbf{Z}_{k+1} - \mathbf{Y}_{k+1} \,|\, \mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\rangle\right].$$

For the scalar product, we have

$$\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\langle \mathbf{Z}_{k+1} - \mathbf{Y}_{k+1} \,|\, \mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\rangle\right]$$
$$= \langle \mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{Z}_{k+1}\right] - \mathbf{Y}_{k+1} \,|\, \mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\rangle,$$

since $\mathbf{Y}_{k+1} \in \mathcal{F}_{k+1/2}$ and $\mathcal{F}_k \subset \mathcal{F}_{k+1/2}$. By (143), the scalar product is zero. Therefore, by **H** 1-*d)* for $\mathbf{Y}_{k+1}$

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Z}_{k+1}\right]\|^2\right]$$
$$\leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right] + \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2\right]$$
$$\leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right] + \sigma_0^2 + \sigma_1^2\mathrm{W}(\boldsymbol{w}_k).$$

By (142), we have

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Z}_{k+1} - \mathbf{Y}_{k+1}\|^2\right] \leq \omega_{\boldsymbol{\mathcal{C}}}\,\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|^2\right].$$

Using (145) with $\zeta_1 = \zeta_2 = 1$ yields our conclusion. □

Proof of Lemma 3. Set $\mathbf{W}_{k+1} := \boldsymbol{H}(\boldsymbol{\mathcal{C}}(\boldsymbol{w}_k, \mathbf{U}_{k+1}), \boldsymbol{X}_{k+1})$ and $\tilde{\boldsymbol{w}}_{k+1} := \boldsymbol{\mathcal{C}}(\boldsymbol{w}_k, \mathbf{U}_{k+1})$. By **CSA** 3 we have that $\boldsymbol{w}_k$ is $\mathcal{F}_k$-measurable, that $\tilde{\boldsymbol{w}}_{k+1}$ is $\mathcal{F}_{k+1/2}$-measurable (as defined in (144)), and that

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\tilde{\boldsymbol{w}}_{k+1} - \boldsymbol{w}_k\|^2\right] = \mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{\mathcal{C}}(\boldsymbol{w}_k, \mathbf{U}_{k+1}) - \boldsymbol{w}_k\|^2\right]$$
$$\leq \kappa_{\boldsymbol{\mathcal{C}}}. \tag{146}$$

We first prove **H** 1-*c)* for $\mathbf{W}_{k+1}$.

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2$$
$$= \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1} - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right] + \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{h}(\tilde{\boldsymbol{w}}_k) - \boldsymbol{h}(\boldsymbol{w}_k)\right]\|^2.$$

Thus for any $\zeta \in \bar{\mathbb{R}}_+$, we get:

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1}\right] - \boldsymbol{h}(\boldsymbol{w}_k)\|^2 \leq (1+\zeta)\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1} - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right]\|^2$$
$$+ (1+\zeta^{-1})\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{h}(\tilde{\boldsymbol{w}}_k) - \boldsymbol{h}(\boldsymbol{w}_k)\right]\|^2.$$

As $\boldsymbol{h}$ is $L_{\boldsymbol{h}}$ Lipschitz, we have an upper bound for the second term: $\boldsymbol{h}(\tilde{\boldsymbol{w}}_k) - \boldsymbol{h}(\boldsymbol{w}_k) \leq L_{\boldsymbol{h}}\|\tilde{\boldsymbol{w}}_k - \boldsymbol{w}_k\|$ almost surely, thus:

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1} - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right]\|^2 \leq L_{\boldsymbol{h}}^2\kappa_{\boldsymbol{\mathcal{C}}}. \tag{147}$$

Moreover, by Jensen inequality,

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{W}_{k+1} - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right]\right]\|^2$$
$$= \|\mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1}) - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right]\right]\|^2$$
$$\leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1}) - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right]\|^2\right]$$

and by **H** 1-*c)* with $\tau_1 = 0$, $\|\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1}) - \boldsymbol{h}(\tilde{\boldsymbol{w}}_k)\right]\|^2 \leq \tau_{0,k} + \tau_{1,k}\mathrm{W}(\tilde{\boldsymbol{w}}_k) \leq \tau_0$. Overall, **H** 1-*c)* is satisfied with $\tau_{0,\boldsymbol{\mathcal{C}}} = (1+\zeta)\tau_0 + (1+\zeta^{-1})L_{\boldsymbol{h}}^2\kappa_{\boldsymbol{\mathcal{C}}}$ and $\tau_{1,\boldsymbol{\mathcal{C}}} = 0$.

We now prove **H** 1-*d)* for $\mathbf{W}_{k+1}$:

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{W}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1}\right]\|^2\right]$$
$$= \mathbb{E}^{\mathcal{F}_k}\left[\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{W}_{k+1} - \mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{W}_{k+1}\right]\|^2\right]\right]$$
$$+ \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{W}_{k+1}\right] - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1}\right]\|^2\right].$$

By **H** 1-*d)*, with $\sigma_1^2 = 0$

$$\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\mathbf{W}_{k+1} - \mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{W}_{k+1}\right]\|^2\right]$$
$$= \mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\|\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1})\right]\|^2\right]$$
$$\leq \sigma_0^2 + \sigma_1^2\mathrm{W}(\tilde{\boldsymbol{w}}_k) \leq \sigma_0^2.$$

Moreover,

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{W}_{k+1}\right] - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1}\right]\|^2\right]$$
$$\leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1})\right] - \mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right]\|^2\right]$$
$$= \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_{k+\frac{1}{2}}}\left[\boldsymbol{H}(\tilde{\boldsymbol{w}}_k, \boldsymbol{X}_{k+1})\right] - \mathbb{E}^{\mathcal{F}_{k+\frac{1}{2}}}\left[\boldsymbol{H}(\boldsymbol{w}_k, \boldsymbol{X}_{k+1})\right]\|^2\right].$$

By assumption, $\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\boldsymbol{H}(\cdot, \boldsymbol{X}_{k+1})\right]$ is $L_{\mathbb{E}\boldsymbol{H}}$-Lipschitz. We conclude using (146):

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_{k+1/2}}\left[\mathbf{W}_{k+1}\right] - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{W}_{k+1}\right]\|^2\right] \leq L_{\mathbb{E}\boldsymbol{H}}^2\kappa_{\boldsymbol{\mathcal{C}}}.$$

We get the result by combining the two bounds above. □

Proof of Lemma 4. Under the constant stepsize assumption, the random field can be written as

$$\widetilde{\boldsymbol{H}}(\boldsymbol{w}_k, \mathbf{U}_{k+1}, \boldsymbol{X}_{k+1}) = \frac{1}{\bar{\gamma}}\left(\boldsymbol{\mathcal{C}}(\boldsymbol{w}_k + \bar{\gamma}\,\mathbf{Y}_{k+1}, \mathbf{U}_{k+1}) - \boldsymbol{w}_k\right).$$

Notice that **H** 1-*c)* can be easily verified since

$$\mathbb{E}^{\mathcal{F}_k}\left[\widetilde{\boldsymbol{H}}(\boldsymbol{w}_k, \mathbf{U}_{k+1}, \boldsymbol{X}_{k+1})\right] = \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right].$$

To verify **H** 1-*d)*, we proceed by

$$\bar{\gamma}^2\mathbb{E}^{\mathcal{F}_k}\left[\|\widetilde{\boldsymbol{H}}(\boldsymbol{w}_k, \mathbf{U}_{k+1}, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|^2\right]$$
$$= \mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{\mathcal{C}}(\boldsymbol{w}_k + \bar{\gamma}\mathbf{Y}_{k+1}, \mathbf{U}_{k+1}) - (\boldsymbol{w}_k + \bar{\gamma}\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right])\|^2\right].$$

As $\boldsymbol{w}_k \in \mathcal{B}_{\boldsymbol{\mathcal{C}}}^d$, applying **CSA** 4 gives

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\widetilde{\boldsymbol{H}}(\boldsymbol{w}_k, \mathbf{U}_{k+1}, \boldsymbol{X}_{k+1}) - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|_2^2\right]$$
$$\leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1} - \mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]\|_2^2\right] + \frac{\Delta_{\boldsymbol{\mathcal{C}}}}{\bar{\gamma}}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|_1\right]$$
$$\leq \sigma_0^2 + \sigma_1^2\mathrm{W}(\boldsymbol{w}_k) + \frac{\Delta_{\boldsymbol{\mathcal{C}}}}{\bar{\gamma}}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|_1\right].$$

Lastly, we obtain the following chain

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathbf{Y}_{k+1}\|_1\right] &\leq \mathbb{E}^{\mathcal{F}_k}\left[\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]-\mathbf{Y}_{k+1}\|_1\right]\\ &\quad \|\boldsymbol{h}(\boldsymbol{w}_k)\|_1+\|\mathbb{E}^{\mathcal{F}_k}\left[\mathbf{Y}_{k+1}\right]-\boldsymbol{h}(\boldsymbol{w}_k)\|_1\\ &\leq \sqrt{d}\left(\sqrt{\tau_{0,k}+\tau_{1,k}\mathrm{W}(\boldsymbol{w}_k)}+\sqrt{c_{h,0}+c_{h,1}\mathrm{W}(\boldsymbol{w}_k)}\right)\\ &\quad +\sqrt{d}\sqrt{\sigma_0^2+\sigma_1^2\mathrm{W}(\boldsymbol{w}_k)}\\ &\leq \sqrt{d}\left(\frac{3+\tau_{0,k}+c_{h,0}+\sigma_0^2}{2}+\frac{\tau_{1,k}+c_{h,1}+\sigma_1^2}{2}\mathrm{W}(\boldsymbol{w}_k)\right),\end{aligned}$$

where we have used $\sqrt{x}\leq\frac{1+x}{2}$, $x\geq 0$ in the last inequality. Collecting terms from the above bounds lead to the desired terms in (62). □

D. Proofs of Section III-B4

Proof of Lemma 9. Let $\boldsymbol{w},\boldsymbol{w}'\in\mathbb{R}^d$. Since a projection is a contraction.

$$\|\mathrm{Prj}_{\varpi}\,\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}}-\mathrm{Prj}_{\varpi}\,\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}'}\|_{\mathbf{D}_{\varpi}}\leq\|\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}}-\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}'}\|_{\mathbf{D}_{\varpi}}.$$

From the expression of B (see (36)), we write

$$\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}}(s)-\mathrm{B}\,\mathcal{V}_{\boldsymbol{w}'}(s)=\lambda\sum_{s'}\mathcal{P}(s,s')\left(\mathcal{V}_{\boldsymbol{w}}(s')-\mathcal{V}_{\boldsymbol{w}'}(s')\right).$$

The squared $\mathbf{D}_{\varpi}$-norm of the RHS is equal to

$$\lambda^2\sum_{s}\varpi(s)\left(\sum_{s'}\mathcal{P}(s,s')\left(\mathcal{V}_{\boldsymbol{w}}(s')-\mathcal{V}_{\boldsymbol{w}'}(s')\right)\right)^2.$$

Since $\mathcal{P}$ is a transition kernel, we have the inequality

$$\left(\sum_{s'}\mathcal{P}(s,s')\left(\mathcal{V}_{\boldsymbol{w}}(s')-\mathcal{V}_{\boldsymbol{w}'}(s')\right)\right)^2\leq\sum_{s'}\mathcal{P}(s,s')\left(\left(\mathcal{V}_{\boldsymbol{w}}(s')-\mathcal{V}_{\boldsymbol{w}'}(s')\right)\right)^2.$$

Finally, since $\varpi\mathcal{P}=\varpi$ by **TD**1, we have

$$\begin{aligned}&\sum_{s}\varpi(s)\sum_{s'}\mathcal{P}(s,s')\left(\left(\mathcal{V}_{\boldsymbol{w}}(s')-\mathcal{V}_{\boldsymbol{w}'}(s')\right)\right)^2\\ &\quad=\sum_{s'}\varpi(s')\left(\left(\mathcal{V}_{\boldsymbol{w}}(s')-\mathcal{V}_{\boldsymbol{w}'}(s')\right)\right)^2\\ &\quad=\|\mathcal{V}_{\boldsymbol{w}}-\mathcal{V}_{\boldsymbol{w}'}\|^2_{\mathbf{D}_{\varpi}}.\end{aligned}$$

This concludes the proof. □

Set

$$\boldsymbol{\mathcal{G}}(s,s'):=\boldsymbol{\phi}(s)\{\lambda\boldsymbol{\phi}(s')-\boldsymbol{\phi}(s)\}^{\top};\tag{148}$$

we may rewrite $\boldsymbol{H}(\boldsymbol{w},(s,s'))$ in (38) as

$$\boldsymbol{H}(\boldsymbol{w},(s,s'))=\boldsymbol{\phi}(s)\mathcal{R}(s,s')+\boldsymbol{\mathcal{G}}(s,s')\boldsymbol{w}.\tag{149}$$

We denote

$$X_0:=(S_0,S_0'),\qquad\bar{\boldsymbol{\mathcal{G}}}_{\varpi}(s,s'):=\boldsymbol{\mathcal{G}}(s,s')-\mathbb{E}_{\varpi}[\boldsymbol{\mathcal{G}}(X_0)].$$

Let $\boldsymbol{w}_\star$ be such that $\mathcal{V}_\star=\boldsymbol{\Phi}\boldsymbol{w}_\star$. Since $\boldsymbol{h}(\boldsymbol{w}_\star)=0$ and $\boldsymbol{h}(\boldsymbol{w})=\mathbb{E}_{\varpi}\left[\boldsymbol{H}(\boldsymbol{w},\boldsymbol{X}_0)\right]$, we get

$$\boldsymbol{h}(\boldsymbol{w})=\boldsymbol{h}(\boldsymbol{w})-\boldsymbol{h}(\boldsymbol{w}_\star)=\mathbb{E}_{\varpi}[\boldsymbol{\mathcal{G}}(X_0)](\boldsymbol{w}-\boldsymbol{w}_\star).\tag{150}$$

We also have

$$\boldsymbol{H}(\boldsymbol{w},(s,s'))-\boldsymbol{H}(\boldsymbol{w}_\star,(s,s'))=\boldsymbol{\mathcal{G}}(s,s')\,(\boldsymbol{w}-\boldsymbol{w}_\star).$$

This yields

$$\begin{aligned}\boldsymbol{H}(\boldsymbol{w},(s,s'))-\boldsymbol{h}(\boldsymbol{w})&=\boldsymbol{H}(\boldsymbol{w}_\star,(s,s'))\\ &\quad+\bar{\boldsymbol{\mathcal{G}}}_{\varpi}(s,s')(\boldsymbol{w}-\boldsymbol{w}_\star).\end{aligned}\tag{151}$$

Set

$$\xi_{\boldsymbol{w}}(s):=(\boldsymbol{w}-\boldsymbol{w}_\star)^{\top}\boldsymbol{\phi}(s)=\mathcal{V}_{\boldsymbol{w}}(s)-\mathcal{V}_{\boldsymbol{w}_\star}(s).\tag{152}$$

From (148) and (150), we have

$$\begin{aligned}\boldsymbol{h}(\boldsymbol{w})&=\mathbb{E}_{\varpi}[\boldsymbol{\mathcal{G}}(S_0,S_0')]\,(\boldsymbol{w}-\boldsymbol{w}_\star)\\ &=\mathbb{E}_{\varpi}[\boldsymbol{\phi}(S_0)\{\lambda\xi_{\boldsymbol{w}}(S_0')-\xi_{\boldsymbol{w}}(S_0)\}].\end{aligned}\tag{153}$$

Proof of Lemma 10. Using (153), we get that

$$\begin{aligned}\|\boldsymbol{h}(\boldsymbol{w})\|^2&=\|\mathbb{E}_{\varpi}[\boldsymbol{\phi}(S_0)\{\lambda\xi_{\boldsymbol{w}}(S_0')-\xi_{\boldsymbol{w}}(S_0)\}]\|^2\\ &\leq\mathbb{E}_{\varpi}[\|\boldsymbol{\phi}(S_0)\|^2]\;\mathbb{E}_{\varpi}[\{\lambda\xi_{\boldsymbol{w}}(S_0')-\xi_{\boldsymbol{w}}(S_0)\}^2].\end{aligned}$$

For the second term, we use

$$\mathbb{E}[(\lambda U+V)^2]\leq\lambda^2\mathbb{E}[U^2]+\mathbb{E}[V^2]+2\lambda\sqrt{\mathbb{E}[U^2]\,\mathbb{E}[V^2]}$$

with $U:=\xi_{\boldsymbol{w}}(S_0')$ and $V:=\xi_{\boldsymbol{w}}(S_0)$. By **TD**1,

$$\mathbb{E}_{\varpi}[\xi^2_{\boldsymbol{w}}(S_0')]=\mathbb{E}_{\varpi}[\xi^2_{\boldsymbol{w}}(S_0)]=\|\mathcal{V}_{\boldsymbol{w}}-\mathcal{V}_\star\|^2_{\mathbf{D}_{\varpi}}$$

which implies that

$$\mathbb{E}_{\varpi}[\{\lambda\xi_{\boldsymbol{w}}(S_0')-\xi_{\boldsymbol{w}}(S_0)\}^2]\leq(1+\lambda)^2\|\mathcal{V}_{\boldsymbol{w}}-\mathcal{V}_\star\|^2_{\mathbf{D}_{\varpi}}.$$

This concludes the proof of the first inequality. From (151) and **TD**3, we obtain

$$\begin{aligned}&\mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})-\boldsymbol{h}(\boldsymbol{w}_k)\|^2\right]\\ &\leq 2\,\mathbb{E}_{\varpi}[\|\boldsymbol{H}(\boldsymbol{w}_\star,X_0)\|^2]\\ &\quad+2\,(\boldsymbol{w}_k-\boldsymbol{w}_\star)^{\top}\mathbb{E}_{\varpi}[\bar{\boldsymbol{\mathcal{G}}}_{\varpi}(X_0)\bar{\boldsymbol{\mathcal{G}}}_{\varpi}^{\top}(X_0)](\boldsymbol{w}_k-\boldsymbol{w}_\star).\end{aligned}$$

We first upper bound $\mathbb{E}_{\varpi}[\|\boldsymbol{H}(\boldsymbol{w}_\star,X_0)\|^2]$. Since $\lambda\leq 1$ and $|\mathcal{R}(s,s')|\leq 1$ (see **TD**2), we get that, for all $\boldsymbol{w}\in\mathbb{R}^d$, and (s,s'), it holds that

$$\begin{aligned}\|\boldsymbol{H}(\boldsymbol{w},(s,s'))\|&\leq 1+\|\boldsymbol{\phi}(s)\{\lambda\mathcal{V}_\star(s')+\mathcal{V}_\star(s)\}\|\\ &\leq 1+\{\lambda|\mathcal{V}_\star(s')|+|\mathcal{V}_\star(s)|\}.\end{aligned}$$

We obtain

$$\mathbb{E}_{\varpi}[\|\boldsymbol{H}(\boldsymbol{w}_\star,X_0)\|^2]\leq 3\left(1+\{\lambda^2+1\}\|\mathcal{V}_\star\|^2_{\mathbf{D}_{\varpi}}\right).$$

Let us upper bound the second term. For any $\mathbf{u}\in\mathbb{R}^d$, it holds that

$$\mathbf{u}^{\top}\mathbb{E}_{\varpi}[\bar{\boldsymbol{\mathcal{G}}}_{\varpi}(X_0)\bar{\boldsymbol{\mathcal{G}}}_{\varpi}^{\top}(X_0)]\mathbf{u}\leq\mathbf{u}^{\top}\mathbb{E}_{\varpi}[\boldsymbol{\mathcal{G}}(X_0)\boldsymbol{\mathcal{G}}^{\top}(X_0)]\mathbf{u},$$

and

$$\mathbf{u}^{T}\boldsymbol{\mathcal{G}}(s,s')\boldsymbol{\mathcal{G}}^{\top}(s,s')\mathbf{u}\leq\{\mathbf{u}^{\top}\boldsymbol{\phi}(s)\}^2(1+\lambda)^2.$$

By combining these two inequalities, we finally obtain

$$\begin{aligned}&(\boldsymbol{w}_k-\boldsymbol{w}_\star)^{\top}\mathbb{E}_{\varpi}[\bar{\boldsymbol{\mathcal{G}}}_{\varpi}(X_0)\bar{\boldsymbol{\mathcal{G}}}_{\varpi}^{\top}(X_0)](\boldsymbol{w}_k-\boldsymbol{w}_\star)\\ &\qquad\leq(1+\lambda)^2\|\mathcal{V}_{\boldsymbol{w}_k}-\mathcal{V}_\star\|^2_{\mathbf{D}_{\varpi}}\end{aligned}$$

This concludes the proof. □

Proof of Lemma 11. It follows from (153) that

$$\begin{aligned}&\langle \boldsymbol{w}-\boldsymbol{w}_* \,|\, \boldsymbol{h}(\boldsymbol{w})-\boldsymbol{h}(\boldsymbol{w}_\star)\rangle\\ &\quad=\mathbb{E}_\varpi[\xi_{\boldsymbol{w}}(S_0)\{\lambda\xi_{\boldsymbol{w}}(S_0')-\xi_{\boldsymbol{w}}(S_0)\}]\\ &\quad=\lambda\mathbb{E}_\varpi[\xi_{\boldsymbol{w}}(S_0)\xi_{\boldsymbol{w}}(S_0')]-\mathbb{E}_\varpi[(\xi_{\boldsymbol{w}}(S_0))^2].\end{aligned}$$

By using the Cauchy-Schwarz inequality, we have

$$\mathbb{E}_\varpi[\xi_{\boldsymbol{w}}(S_0)\xi_{\boldsymbol{w}}(S_0')]\le\{\mathbb{E}_\varpi[\xi^2_{\boldsymbol{w}}(S_0)]\}^{1/2}\{\mathbb{E}_\varpi[\xi^2_{\boldsymbol{w}}(S_0')]\}^{1/2}.$$

Under **TD**1, $\varpi\mathcal{P}=\varpi$, which implies that for any function g,

$$\mathbb{E}_\varpi[g(S_0')]=\mathbb{E}_\varpi[g(S_0)].$$

This yields

$$\langle \boldsymbol{w}-\boldsymbol{w}_* \,|\, \boldsymbol{h}(\boldsymbol{w})-\boldsymbol{h}(\boldsymbol{w}_\star)\rangle\le-(1-\lambda)\mathbb{E}_\varpi[\xi^2_{\boldsymbol{w}}(S_0)].$$

The proof is concluded by using (152) and

$$\begin{aligned}\mathbb{E}_\varpi[\xi^2_{\boldsymbol{w}}(S_0)]&=\mathbb{E}_\varpi[(\mathcal{V}_{\boldsymbol{w}}(S_0)-\mathcal{V}_\star(S_0))^2]\\&=\|\mathcal{V}_{\boldsymbol{w}}-\mathcal{V}_\star\|^2_{\mathbf{D}_\varpi}.\end{aligned}$$

□

Proof of Lemma 12. Note first that

$$\begin{aligned}\boldsymbol{w}^\top\boldsymbol{\Phi}^\top\mathbf{D}_\varpi\boldsymbol{\Phi}\boldsymbol{w}&=\sum_{s\in\mathcal{S}}\varpi(s)\langle\boldsymbol{\phi}(s)\,|\,\boldsymbol{w}\rangle^2\\&\le\sum_{s\in\mathcal{S}}\varpi(s)\|\boldsymbol{\phi}(s)\|^2\|\boldsymbol{w}\|^2\le\|\boldsymbol{w}\|^2,\end{aligned}$$

where we have used that $\sum_{s\in\mathcal{S}}\varpi(s)=1$. On the other hand, under **TD** 1, $\mathbf{D}_\varpi$ is full rank ($\mathbf{D}_\varpi$ is diagonal and all its diagonal entries are positive), which implies that the minimal eigenvalue of $\boldsymbol{\Phi}^\top\mathbf{D}_\varpi\boldsymbol{\Phi}$ is positive. The result follows. □

E. Proofs of Section IV-A

We hereafter give a slightly different statement of Lemma 13 for the particular case in which $\nabla V=-\boldsymbol{h}$.

H 3. *The field h and the function V in **H** 2 satisfy: $\nabla\mathrm{V}=-\boldsymbol{h}$.*

Then under **H** 2-*c*), the Lyapunov functions V, W satisfy, for any $\boldsymbol{w}\in\mathbb{R}^d$,

$$-\|\nabla\mathrm{V}(\boldsymbol{w})\|^2=\langle\nabla\mathrm{V}(\boldsymbol{w})\,|\,\boldsymbol{h}(\boldsymbol{w})\rangle\le-\varrho\mathrm{W}(\boldsymbol{w})\,. \tag{154}$$

Define, for any $k\ge0$:

$$\omega_{2,k+1}:=(\varrho-\tau_1)-\gamma_{k+1}L_{\mathrm{V}}\sigma_1^2 \tag{155}$$

$$\gamma_{2,\max}:=\min\left(\frac{1}{L_{\mathrm{V}}},\frac{(\varrho-\tau_1)}{L_{\mathrm{V}}\sigma_1^2}\right). \tag{156}$$

Lemma 18 (Robbins-Siegmund type inequality for $\nabla V=-\boldsymbol{h}$)**.** *Assume **H** 1 and 2, **NA** 1 and **H** 3. Then, for any $k\ge0$, we have almost-surely*

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right]&\le\mathrm{V}(\boldsymbol{w}_k)-\frac{\gamma_{k+1}}{2}\omega_{2,k+1}\mathrm{W}(\boldsymbol{w}_k)\\&\quad+\frac{\gamma_{k+1}}{2}\tau_0+\frac{\gamma_{k+1}^2L_{\mathrm{V}}}{2}\sigma_0^2,\end{aligned} \tag{157}$$

Proof. Let $k\ge0$. By **H** 2-*b*), we have

$$\begin{aligned}\mathrm{V}(\boldsymbol{w}_{k+1})\le\mathrm{V}(\boldsymbol{w}_k)+\langle\nabla\mathrm{V}(\boldsymbol{w}_k)\,|\,\boldsymbol{w}_{k+1}-\boldsymbol{w}_k\rangle\\+(L_{\mathrm{V}}/2)\|\boldsymbol{w}_{k+1}-\boldsymbol{w}_k\|^2\,.\end{aligned}$$

Computing the conditional expectation of both sides of this inequality yields

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right]&\le\mathrm{V}(\boldsymbol{w}_k)\\&\quad+\gamma_{k+1}\langle\nabla\mathrm{V}(\boldsymbol{w}_k)\,|\,\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\rangle+\\&\quad+\gamma_{k+1}^2(L_{\mathrm{V}}/2)\,\mathbb{E}^{\mathcal{F}_k}[\|\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\|^2]\,.\end{aligned}$$

We now use $2\langle a\,|\,b\rangle=\|a\|^2+\|b\|^2-\|a-b\|^2$.

$$\begin{aligned}&2\langle\nabla\mathrm{V}(\boldsymbol{w}_k)\,|\,\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\rangle\\&=-2\langle-\nabla\mathrm{V}(\boldsymbol{w}_k)\,|\,\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\rangle\\&=-\|-\nabla\mathrm{V}(\boldsymbol{w}_k)\|^2-\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\|^2\\&\quad+\|-\nabla\mathrm{V}(\boldsymbol{w}_k)-\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\|^2.\end{aligned}$$

Define $\boldsymbol{b}_k:=\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]-\boldsymbol{h}(\boldsymbol{w}_k)$. We have, using Equation (154),

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right]&\le\mathrm{V}(\boldsymbol{w}_k)-\frac{\varrho\gamma_{k+1}}{2}\mathrm{W}(\boldsymbol{w}_k)\\&\quad-\frac{\gamma_{k+1}}{2}\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\|^2+\frac{\gamma_{k+1}}{2}\|\boldsymbol{b}_k\|^2\\&\quad+\gamma_{k+1}^2(L_{\mathrm{V}}/2)\,\mathbb{E}^{\mathcal{F}_k}[\|\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\|^2]\,.\end{aligned}$$

Note first that, using **H** 1-*c*) we get

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right]&\le\mathrm{V}(\boldsymbol{w}_k)-\frac{\varrho\gamma_{k+1}}{2}\mathrm{W}(\boldsymbol{w}_k)\\&\quad-\frac{\gamma_{k+1}}{2}\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\|^2+\frac{\gamma_{k+1}}{2}(\tau_0+\tau_1\mathrm{W}(\boldsymbol{w}_k))\\&\quad+\gamma_{k+1}^2(L_{\mathrm{V}}/2)\,\mathbb{E}^{\mathcal{F}_k}[\|\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\|^2]\,.\end{aligned}$$

We compute a bias-variance decomposition and use **H** 1-*d*):

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}[\|\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\|^2]&=\|\mathbb{E}^{\mathcal{F}_k}[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})]\|^2\\&\quad+\mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})-\mathbb{E}^{\mathcal{F}_k}[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})]\|^2\right]\\&\le\|\mathbb{E}^{\mathcal{F}_k}[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})]\|^2+\sigma_0^2+\sigma_1^2\mathrm{W}(\boldsymbol{w}_k).\end{aligned}$$

Overall, we get:

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right]&\le\mathrm{V}(\boldsymbol{w}_k)\\&\quad-\left(\frac{\gamma_{k+1}}{2}(\varrho-\tau_1)-\frac{\gamma_{k+1}^2L_{\mathrm{V}}}{2}\sigma_1^2\right)\mathrm{W}(\boldsymbol{w}_k)\\&\quad-\left(\frac{\gamma_{k+1}}{2}-\frac{\gamma_{k+1}^2L_{\mathrm{V}}}{2}\right)\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]\|^2\\&\quad+\frac{\gamma_{k+1}}{2}\tau_0+\frac{\gamma_{k+1}^2L_{\mathrm{V}}}{2}\sigma_0^2.\end{aligned}$$

We thus get:

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right]&\le\mathrm{V}(\boldsymbol{w}_k)-\frac{\gamma_{k+1}}{2}\omega_{2,k+1}\mathrm{W}(\boldsymbol{w}_k)\\&\quad+\frac{\gamma_{k+1}}{2}\tau_0+\frac{\gamma_{k+1}^2L_{\mathrm{V}}}{2}\sigma_0^2,\end{aligned}$$

with $\omega_{2,k}$, $\gamma_{2,\max}$ as in (155), (156). □

F. Proofs of Section IV-B2

Proof of Proposition 2. By Corollary 1, the field $\boldsymbol{H}(\boldsymbol{w},\sim)=\mathrm{Top}_1(\nabla\mathrm{F}(\mathrm{w}))$ satisfies **H** 1 with $(c_{\boldsymbol{h},0},c_{\boldsymbol{h},1})=(0,1)$, $(\tau_{0,k},\tau_{1,k})=(0,1-1/d)$ and $(\sigma_0^2,\sigma_1^2)=(0,0)$, for $\mathrm{W}=\|\nabla F(\cdot)\|^2$).

We consider $\mathrm{V} = F$ and thus have **H** 2 with $c_{\mathrm{V}} = 1$, $\rho = 1$, $L_{\mathrm{V}} = L_{\nabla F}$. We thus verify Equations (73) to (76) with:

$$\begin{aligned}
\mathrm{b}_0 &:= c_{\mathrm{V}}\sqrt{\tau_0}/2 = 0\\
\mathrm{b}_1 &:= c_{\mathrm{V}}(\sqrt{\tau_0}/2 + \sqrt{\tau_1}) = \sqrt{1 - 1/d} \leq 1 - \frac{1}{2d}.\\
\eta_0 &:= 0\\
\eta_1 &:= \sigma_1^2 + \tau_1 + c_{\boldsymbol{h},1} + \sqrt{c_{\boldsymbol{h},1}}\,(\sqrt{\tau_0} + \sqrt{\tau_1})\\
&\qquad + \sqrt{\tau_1}\left(\sqrt{c_{\boldsymbol{h},0}} + \sqrt{c_{\boldsymbol{h},1}}\right) \leq 4,\\
\gamma_{\max} &:= 2\{\varrho - \mathrm{b}_1\}/(L_{\mathrm{V}}\eta_1) \geq \frac{1}{dL_{\mathrm{V}}\eta_1}\,,\\
\omega_k &:= 2\{\varrho - \mathrm{b}_1\} - \gamma_k L_{\mathrm{V}}\eta_1 \geq \frac{1}{2d}\,.
\end{aligned}$$

Proposition 2 is then a direct application of Theorem 1 with the constants above. □

For the proofs of the next two propositions, we note that the random field $\boldsymbol{H}$ satisfies **H** 1 with $\mathrm{V} = F$, $\mathrm{W} = \|\nabla F(\cdot)\|^2$. Moreover, the constants are $(c_{\boldsymbol{h},0}, c_{\boldsymbol{h},1}) = (0,1)$, $(\tau_{0,k}, \tau_{1,k}) = (0,0)$, $(\sigma_0^2, \sigma_1^2) = (M^2/n, 0)$, and $\varrho = 1$.

Proof of Proposition 3. By Lemma 2, the compressed SG in (16) uses a random field that satisfies **H** 1 with the same set of constants that inherit from $\boldsymbol{H}$ except for $\sigma_{0;\boldsymbol{\mathcal{C}}}^2 := (1 + \omega_{\boldsymbol{\mathcal{C}}})\sigma_0^2$, $\sigma_{1;\boldsymbol{\mathcal{C}}}^2 := 2\omega_{\boldsymbol{\mathcal{C}}}$. Thus we have (73) with $\mathrm{b}_0 = \mathrm{b}_1 = 0$ and (74), (75), (76) with

$$\begin{aligned}
\eta_0 &:= (1 + \omega_{\boldsymbol{\mathcal{C}}})M^2/n, \qquad \eta_1 := 2\omega_{\boldsymbol{\mathcal{C}}} + 1,\\
\gamma_{\max} &:= 2/(L_{\nabla F}\eta_1) = 2/(L_{\nabla F}(2\omega_{\boldsymbol{\mathcal{C}}} + 1))\,,\\
\omega_k &:= 2 - \gamma_k L_{\nabla F}\eta_1 \geq 1\,.
\end{aligned}$$

Where the last equation holds as $\gamma_k \leq \gamma_{\max}/2$. As before, Proposition 3 is then a direct application of Theorem 1 with the constants above. □

Proof of Proposition 4. For $\boldsymbol{\mathcal{C}}$ is Q_d, **CSA** 4 is satisfied with $\kappa_{\boldsymbol{\mathcal{C}}} = d\Delta^2$. Using Corollary 2, we can apply the result of Theorem 1, with $\tau_{0,k,\boldsymbol{\mathcal{C}}} := L_{\nabla F}^2 d\Delta^2$, $\sigma_{0;\boldsymbol{\mathcal{C}}}^2 := \frac{M^2}{n} + L_{\nabla F}^2 d\Delta^2$. Thus

$$\begin{aligned}
\mathrm{b}_0 &= L_{\nabla F}\sqrt{d}\Delta/2 \leq 1/2\,, \quad \mathrm{b}_1 = L_{\nabla F}\sqrt{d}\Delta/2 \leq 1/2,\\
\eta_0 &= \frac{M^2}{n} + 2\,L_{\nabla F}^2 d\Delta^2 + L_{\nabla F}\sqrt{d}\Delta \leq \frac{M^2}{n} + 3L_{\nabla F}\sqrt{d}\Delta\\
\eta_1 &:= 1 + \left(L_{\nabla F}\sqrt{d}\Delta\right) \leq 2\\
\gamma_{\max} &:= 2\{1 - \mathrm{b}_1\}/(L_{\nabla F}\eta_1) \geq 1/(2L_{\nabla F})\\
\omega_k &:= 2\{\varrho - \mathrm{b}_1\} - \gamma_k L_{\mathrm{V}}\eta_1 \geq 1/2\,.
\end{aligned}$$

Proposition 4 is then a direct application of Theorem 1 with the constants above. □

Proof of Proposition 5. By Lemma 4, the compressed SG in (18) uses a random field (19) that satisfies **H** 1 with the same set of constants inherited from $\boldsymbol{H}$ except for

$$\sigma_{0;\boldsymbol{\mathcal{C}}}^2 := \frac{M^2}{n} + \frac{\Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d}}{2\bar{\gamma}}(3 + M^2/n), \quad \sigma_{1;\boldsymbol{\mathcal{C}}}^2 := \frac{\Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d}}{2\bar{\gamma}}.$$

We also have (73) with $\mathrm{b}_0 = \mathrm{b}_1 = 0$ and (74), (75) with

$$\eta_0 := M^2/n + (M^2/n)\frac{\Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d}}{2\bar{\gamma}}, \quad \eta_1 := 1 + \frac{\Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d}}{2\bar{\gamma}},$$

$$\gamma_{\max} := 2/(L_{\nabla F}(1 + \Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d}/(2\bar{\gamma}))\,.$$

To apply Theorem 1, we have to satisfy $\bar{\gamma} < \gamma_{\max}$, which is then equivalent to

$$\bar{\gamma} < \frac{2}{L_{\nabla F}(1 + \frac{\Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d}}{2\bar{\gamma}})} \iff \bar{\gamma} + \Delta_{\boldsymbol{\mathcal{C}}}\sqrt{d} < \frac{2}{L_{\nabla F}}.$$

This yields the stepsize condition in (102). Our bound in (103) is then achieved by plugging the above η_0 into (82). □

G. Proofs of Section IV-B3c

From Section III-B3b, it holds

$$\begin{aligned}
\eta_0 &= 12s_\star\sqrt{c_{\chi,0}}/\mathsf{m} + 144s_\star^2 c_{\chi,0}/\mathsf{m}^2 + 4s_\star^2\sqrt{c_{\chi,0} + c_{\chi,1}}/(n\mathsf{m}),\\
\eta_1 &= 1 + 12s_\star\left(\sqrt{c_{\chi,0}} + \sqrt{c_{\chi,1}}\right)/\mathsf{m} + 144s_\star^2 c_{\chi,1}/\mathsf{m}^2\\
&\qquad + 4s_\star^2\sqrt{c_{\chi,1}}/(n\mathsf{m}),\\
\mathrm{b}_0 &= 6s_\star\sqrt{v_{\max}}\sqrt{c_{\chi,0}}/\mathsf{m},\\
\mathrm{b}_1 &= 6s_\star\sqrt{v_{\max}}\left(\sqrt{c_{\chi,0}} + 2\sqrt{c_{\chi,1}}\right)/\mathsf{m},\\
\gamma_{\max} &= 2\{v_{\min} - \mathrm{b}_1\}/(L_{\mathrm{V}}\eta_1),\\
\omega_k &= 2\{v_{\min} - \mathrm{b}_1\} - \gamma_k L_{\mathrm{V}}\eta_1.
\end{aligned}$$

H. Proofs of Section V

Proof of Theorem 4. We first establish that the sequence $\{\boldsymbol{w}_k,\ k \in \mathbb{N}\}$ satisfies **SA** 3 with probability 1. Define

$$\mathrm{b}_{0,k} := c_{\mathrm{V}}\sqrt{\tau_{0,k}}/2, \quad \mathrm{b}_{1,k} := c_{\mathrm{V}}(\sqrt{\tau_{0,k}}/2 + \sqrt{\tau_{1,k}}).$$

It follows from the Robbins-Siegmund inequality (see Lemma 13) that

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right] &\leq \mathrm{V}(\boldsymbol{w}_k) + \gamma_{k+1}\,\mathrm{b}_{0,k} + \gamma_{k+1}^2 L_{\mathrm{V}}\eta_0/2\\
&\quad - \gamma_{k+1}\{\varrho - \mathrm{b}_{1,k} - \gamma_{k+1}L_{\mathrm{V}}\,\eta_1/2\}\,\mathrm{W}(\boldsymbol{w}_k).
\end{aligned}$$

By $\sum_{k=0}^{\infty}\gamma_k = \infty$ and $\sum_{k=0}^{\infty}\gamma_k^2 < \infty$, and i),

$$\lim_{k\to\infty}\{\varrho - \mathrm{b}_{1,k} - \gamma_{k+1}L_{\mathrm{V}}\,\eta_1/2\} = \varrho \tag{158}$$

exists and is positive. Therefore, there exists k_0 such that for all $k \geq k_0$,

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right] &\leq \mathrm{V}(\boldsymbol{w}_k) + \gamma_{k+1}\,\mathrm{b}_{0,k} + \gamma_{k+1}^2 L_{\mathrm{V}}\eta_0/2\\
&\quad - \gamma_{k+1}\{\varrho - \mathrm{b}_{1,k} - \gamma_{k+1}L_{\mathrm{V}}\,\eta_1/2\}\,(\mathrm{W}(\boldsymbol{w}_k) \wedge C)
\end{aligned} \tag{159}$$

where C is a positive constant, and

$$\mathbb{E}^{\mathcal{F}_k}\left[\mathrm{V}(\boldsymbol{w}_{k+1})\right] \leq \mathrm{V}(\boldsymbol{w}_k) + \gamma_{k+1}\,\mathrm{b}_{0,k} + \gamma_{k+1}^2 L_{\mathrm{V}}\eta_0/2\,. \tag{160}$$

Define for $k \geq k_0$,

$$M_k := \mathrm{V}(\boldsymbol{w}_k) - \mathrm{V}_\star + \sum_{\ell=k+1}^{\infty}\mathrm{c}_\ell, \tag{161}$$

where $\mathrm{c}_{\ell+1} := \gamma_{\ell+1}\,\mathrm{b}_{0,\ell} + (L_{\mathrm{V}}\eta_0/2)\gamma_\ell^2$. It is easily checked that for all $k \geq k_0$, $M_k \geq 0$ and $\mathbb{E}^{\mathcal{F}_k}\left[M_{k+1}\right] \leq M_k$ almost-surely. Hence, $\{M_{\ell-k_0},\ \ell - k_0 \in \mathbb{N}\}$ is a non-negative super-martingale. It follows from [189, Theorems II-2-7,II-2-9] that $\sup_{k\in\mathbb{N}} M_k < \infty$ and the sequence $\{M_k,\ k \in \mathbb{N}\}$ converges with probability one. This implies that, with probability one, $\sup_k \mathrm{V}(\boldsymbol{w}_k) < \infty$ and $\lim_{k\to\infty}\mathrm{V}(\boldsymbol{w}_k)$ exists. By ii), this

yields $\sup_{k\in\mathbb{N}}\|\boldsymbol{w}_k\|<\infty$ with probability one. Hence, **SA** 3 holds with probability one.

Note two corollaries of the discussion above. First, from (158), (159), $\sup_k \mathrm{V}(\boldsymbol{w}_k)<\infty$, the stepsize condition $\sum_{k=0}^{\infty}\gamma_k=\infty$, $\sum_{k=0}^{\infty}\gamma_k^2<\infty$, and i), we have when $C\to+\infty$,

$$\sup_k \mathrm{W}(\boldsymbol{w}_k)<\infty$$

with probability one. Second, it also follows from (160) that

$$\sup_{k\in\mathbb{N}}\mathbb{E}[\mathrm{V}(\boldsymbol{w}_k)]\le\mathbb{E}[\mathrm{V}(\boldsymbol{w}_0)]+\sum_{k=0}^{\infty}\gamma_{k+1}\{\mathrm{b}_{0,k}+(L_{\mathrm{V}}\sigma_0^2/2)\gamma_{k+1}\};$$

the RHS is finite by the stepsize conditions, i) and the condition $\mathbb{E}[\mathrm{V}(\boldsymbol{w}_0)]<\infty$.

We now consider **SA** 4. We proved that $\sup_k \mathrm{W}(\boldsymbol{w}_k)<\infty$ with probability one. Combined with i), this yields with probability one,

$$\|\mathbb{E}^{\mathcal{F}_k}\left[\boldsymbol{H}(\boldsymbol{w}_k,\boldsymbol{X}_{k+1})\right]-\boldsymbol{h}(\boldsymbol{w}_k)\|^2 \le \tau_{0,k}+\tau_{1,k}\mathrm{W}(\boldsymbol{w}_k)\to 0.$$

The asymptotic rate of change condition is a consequence of the convergence theorem for square integrable martingales (see e.g. [190, Theorem 2.15]): the series $\sum_{k=1}^{\infty}\gamma_k\boldsymbol{u}_{k+1}$ converges with probability one on the event $\sum_{k=1}^{\infty}\gamma_k^2\mathbb{E}^{\mathcal{F}_k}\left[\|\boldsymbol{u}_{k+1}\|^2\right]<\infty$. We thus have to prove that $\sum_k\gamma_{k+1}^2\sigma_0^2+\sum_k\gamma_{k+1}^2\sigma_1^2\mathrm{W}(\boldsymbol{w}_k)<\infty$ with probability one. This holds true by the stepsize conditions and the property $\sup_k \mathrm{W}(\boldsymbol{w}_k)<\infty$. □

I. Proofs of Section VI

Proof of Lemma 14. Let $t\in\{1,\dots,\mathrm{k}_{\mathrm{out}}\}$ and $k\in\{0,\dots,\mathrm{k}_{\mathrm{in}}-1\}$. Since $\mathrm{b}_{\mathrm{vr}}^{-1}\mathbb{E}\left[\sum_{i\in\mathcal{B}_{t,k+1}}a_i\right]=n^{-1}\sum_{i=1}^n a_i$ (see e.g. [38, Lemma 7.1.]) and $\boldsymbol{H}_{t,k}^{\mathrm{vr}}\in\mathcal{F}_{t,k}$, then

$$\mathbb{E}^{\mathcal{F}_{t,k}}\left[\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}\right]=\boldsymbol{H}_{t,k}^{\mathrm{vr}}+n^{-1}\sum_{i=1}^n\left(\boldsymbol{h}_i(\boldsymbol{w}_{t,k})-\boldsymbol{h}_i(\boldsymbol{w}_{t,k-1})\right).$$

This concludes the proof of the first claim. The second one follows by induction, upon noting that $\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathbf{U}\right]=\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathbb{E}^{\mathcal{F}_{t,\ell}}\left[\mathbf{U}\right]\right]$ for any $\ell\ge 0$. □

Proof of Lemma 15. Let $t\in\{1,\dots,\mathrm{k}_{\mathrm{out}}\}$ and $k\in\{0,\dots,\mathrm{k}_{\mathrm{in}}-1\}$. By Lemma 14, we have

$$\begin{aligned}&\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\mathbb{E}^{\mathcal{F}_{t,k}}\left[\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}\right]\\&=\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\boldsymbol{h}(\boldsymbol{w}_{t,k})-\boldsymbol{H}_{t,k}^{\mathrm{vr}}+\boldsymbol{h}(\boldsymbol{w}_{t,k-1})\\&=\frac{1}{\mathrm{b}_{\mathrm{vr}}}\sum_{i\in\mathcal{B}_{k+1}}\Delta_i(\boldsymbol{w}_{t,k},\boldsymbol{w}_{t,k-1})-\frac{1}{n}\sum_{i=1}^n\Delta_i(\boldsymbol{w}_{t,k},\boldsymbol{w}_{t,k-1})\end{aligned}$$

where $\Delta_i(\boldsymbol{w}_{t,k},\boldsymbol{w}_{t,k-1}):=\boldsymbol{h}_i(\boldsymbol{w}_{t,k})-\boldsymbol{h}_i(\boldsymbol{w}_{t,k-1})$. The conditional expectation of the L^2-moment is upper bounded by $\mathrm{b}_{\mathrm{vr}}^{-1}n^{-1}\sum_{i=1}^n\|\Delta_i(\boldsymbol{w}_{t,k},\boldsymbol{w}_{t,k-1})\|^2$ (see e.g. [38, Lemma 7.1.]). Under **VR** 1, we have $\|\Delta_i(\boldsymbol{w}_{t,k},\boldsymbol{w}_{t,k-1})\|^2\le L_i^2\|\boldsymbol{w}_{t,k}-\boldsymbol{w}_{t,k-1}\|^2$. This concludes the proof of the first claim. For the second one, we write $\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\boldsymbol{h}(\boldsymbol{w}_{t,k})=\mathbf{U_1}+\mathbf{U_2}$ where $\mathbf{U_1}:=\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\mathbb{E}^{\mathcal{F}_{t,k}}\left[\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}\right]$. By definition of the conditional expectation and since $\mathbf{U_2}\in\mathcal{F}_{t,k}$, we have $\mathbb{E}^{\mathcal{F}_{t,k}}\left[\|\mathbf{U_1}+\mathbf{U_2}\|^2\right]=\mathbb{E}^{\mathcal{F}_{t,k}}\left[\|\mathbf{U_1}\|^2\right]+\mathbb{E}^{\mathcal{F}_{t,k}}\left[\|\mathbf{U_2}\|^2\right]=\mathbb{E}^{\mathcal{F}_{t,k}}\left[\|\mathbf{U_1}\|^2\right]+\|\mathbf{U_2}\|^2$. The proof is concluded by using the first statement and Lemma 14. □

Proof of Lemma 16. From **H** 2-*b*), we write for any $\boldsymbol{w},\boldsymbol{d},\boldsymbol{h}\in\mathbb{R}^d$ and $\gamma,\beta>0$,

$$\begin{aligned}\mathrm{V}(\boldsymbol{w}+\gamma\boldsymbol{d})&\le\mathrm{V}(\boldsymbol{w})+\gamma\langle\nabla\mathrm{V}(\boldsymbol{w})\,|\,\boldsymbol{d}\rangle+\gamma^2\frac{L_{\mathrm{V}}}{2}\|\boldsymbol{d}\|^2\\&\le\mathrm{V}(\boldsymbol{w})+\gamma\langle\nabla\mathrm{V}(\boldsymbol{w})\,|\,\boldsymbol{h}\rangle+\gamma^2L_{\mathrm{V}}\left(\|\boldsymbol{d}-\boldsymbol{h}\|^2+\|\boldsymbol{h}\|^2\right)\\&\quad+\frac{\gamma}{2\beta}\|\nabla\mathrm{V}(\boldsymbol{w})\|^2+\frac{\gamma\beta}{2}\|\boldsymbol{d}-\boldsymbol{h}\|^2.\end{aligned}$$

Applied with $\boldsymbol{h}\leftarrow\boldsymbol{h}(\boldsymbol{w})$, and using **H** 1-*b*), **H** 2-*c*) and (46), this inequality implies

$$\begin{aligned}\mathrm{V}(\boldsymbol{w}+\gamma\boldsymbol{d})&\le\mathrm{V}(\boldsymbol{w})-\gamma\left(\varrho-\frac{1}{2\beta}c_{\mathrm{V}}^2-\gamma L_{\mathrm{V}}c_{\boldsymbol{h},1}\right)\mathrm{W}(\boldsymbol{w})\\&\quad+\gamma\left(\frac{\beta}{2}+\gamma L_{\mathrm{V}}\right)\|\boldsymbol{d}-\boldsymbol{h}(\boldsymbol{w})\|^2+\gamma^2L_{\mathrm{V}}c_{\boldsymbol{h},0}.\end{aligned}$$

We choose $\beta:=c_{\mathrm{V}}^2/\varrho$ and obtain, for any $\mu>0$,

$$\begin{aligned}\mathrm{V}(\boldsymbol{w}+\gamma\boldsymbol{d})&\le\mathrm{V}(\boldsymbol{w})-\gamma\left(\frac{\varrho}{2}-\gamma L_{\mathrm{V}}c_{\boldsymbol{h},1}\right)\mathrm{W}(\boldsymbol{w})+\gamma^2L_{\mathrm{V}}c_{\boldsymbol{h},0}\\&\quad-\gamma\mu\|\boldsymbol{d}-\boldsymbol{h}(\boldsymbol{w})\|^2+\gamma\left(\frac{c_{\mathrm{V}}^2}{2\varrho}+\mu+\gamma L_{\mathrm{V}}\right)\|\boldsymbol{d}-\boldsymbol{h}(\boldsymbol{w})\|^2.\end{aligned}$$

Applying this inequality with $\gamma=\gamma_{t,k+1}$, $\boldsymbol{w}=\boldsymbol{w}_{t,k}$, $\boldsymbol{d}=\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}$, $\mu=\mu_{t,k+1}:=\varrho/2-\gamma_{t,k+1}L_{\mathrm{V}}$ and computing the conditional expectation *w.r.t.* to $\mathcal{F}_{t,0}$, concludes the proof. □

Proof of Theorem 7. By Equation (134) and **H** 1-*b*)

$$\begin{aligned}&\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right]\\&\quad\le 2\frac{L^2\mathrm{k}_{\mathrm{in}}}{\mathrm{b}_{\mathrm{vr}}}\sum_{k=1}^{\mathrm{k}_{\mathrm{in}}}\gamma_{t,k}^3\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}_{t,k}^{\mathrm{vr}}-\boldsymbol{h}(\boldsymbol{w}_{t,k-1})\|^2\right]\\&\quad+2\frac{L^2\mathrm{k}_{\mathrm{in}}}{\mathrm{b}_{\mathrm{vr}}}\sum_{k=1}^{\mathrm{k}_{\mathrm{in}}}\gamma_{t,k}^3(c_{\boldsymbol{h},0}+c_{\boldsymbol{h},1}\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathrm{W}(\boldsymbol{w}_{t,k-1})\right]).\end{aligned}$$

We use Lemma 16 and sum from $k=0$ to $k=\mathrm{k}_{\mathrm{in}}-1$:

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathrm{V}(\boldsymbol{w}_{t,\mathrm{k}_{\mathrm{in}}})\right]&\le\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathrm{V}(\boldsymbol{w}_{t,0})\right]+L_{\mathrm{V}}c_{\boldsymbol{h},0}\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}^2\\&-\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}\nu_{t,k+1}\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathrm{W}(\boldsymbol{w}_{t,k})\right]\\&-\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}\,\mu_{t,k+1}\,\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right]\\&+2\frac{L^2\mathrm{k}_{\mathrm{in}}}{\mathrm{b}_{\mathrm{vr}}}\,\mathrm{a}\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}^3\mathbb{E}^{\mathcal{F}_{t,0}}\left[\|\boldsymbol{H}_{t,k+1}^{\mathrm{vr}}-\boldsymbol{h}(\boldsymbol{w}_{t,k})\|^2\right]\\&+2\frac{L^2\mathrm{k}_{\mathrm{in}}}{\mathrm{b}_{\mathrm{vr}}}c_{\boldsymbol{h},0}\,\mathrm{a}\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}^3\\&+2\frac{L^2\mathrm{k}_{\mathrm{in}}}{\mathrm{b}_{\mathrm{vr}}}c_{\boldsymbol{h},1}\,\mathrm{a}\sum_{k=0}^{\mathrm{k}_{\mathrm{in}}-1}\gamma_{t,k+1}^3\mathbb{E}^{\mathcal{F}_{t,0}}\left[\mathrm{W}(\boldsymbol{w}_{t,k})\right].\end{aligned}$$

Hence, we get

$$\begin{aligned}
&\sum_{k=1}^{\mathrm{k_{in}}} \gamma_{t,k} \left(\frac{\varrho}{2} - c_{\boldsymbol{h},1}\, \gamma_{t,k}\, \lambda_{t,k} \right) \mathbb{E}^{\mathcal{F}_{t,0}} \left[\mathrm{W}(\boldsymbol{w}_{t,k-1}) \right] \\
&\quad + \sum_{k=1}^{\mathrm{k_{in}}} \gamma_{t,k} \left(\frac{\varrho}{2} - \gamma_{t,k}\, \lambda_{t,k} \right) \mathbb{E}^{\mathcal{F}_{t,0}} \left[\| \boldsymbol{H}_{t,k}^{\mathrm{vr}} - \boldsymbol{h}(\boldsymbol{w}_{t,k-1}) \|^2 \right] \\
&\leq \mathbb{E}^{\mathcal{F}_{t,0}} \left[\mathrm{V}(\boldsymbol{w}_{t,0}) \right] - \mathbb{E}^{\mathcal{F}_{t,0}} \left[\mathrm{V}(\boldsymbol{w}_{t,\mathrm{k_{in}}}) \right] \\
&\quad + c_{\boldsymbol{h},0} \left\{ L_{\mathrm{V}} \sum_{k=1}^{\mathrm{k_{in}}} \gamma_{t,k}^2 + 2 \frac{L^2 \mathrm{k_{in}}}{\mathsf{b}_{\mathrm{vr}}}\, \mathrm{a} \sum_{k=1}^{\mathrm{k_{in}}} \gamma_{t,k}^3 \right\}.
\end{aligned}$$

We sum from $t = 1$ to $t = \mathrm{k_{out}}$, and use **H** 2-*a*), $\boldsymbol{w}_{t-1,\mathrm{k_{in}}} = \boldsymbol{w}_{t,0}$ (see Line 3 in Algorithm 1) to conclude the proof. □